Gregor Betz

# Prediction or Prophecy?

## The Boundaries of Economic Foreknowledge and Their Socio-Political Consequences

DUV

Gregor Betz

# Prediction or Prophecy?

WIRTSCHAFTSWISSENSCHAFT

Gregor Betz

# Prediction or Prophecy?

## The Boundaries of Economic Foreknowledge and Their Socio-Political Consequences

With a foreword by Prof. Dr. Holm Tetens

Deutscher Universitäts-Verlag

Bibliografische Information Der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Dissertation Freie Universität Berlin, 2004

Gedruckt mit Unterstützung des Förderungs- und Beihilfefonds Wissenschaft der VG WORT

1. Auflage September 2006



Lektorat: Ute Wrasmann / Anita Wilke

Der Deutsche Universitäts-Verlag ist ein Unternehmen von Springer Science+Business Media.
www.duv.de



Umschlaggestaltung: Regine Zimmer, Dipl.-Designerin, Frankfurt/Main
Druck und Buchbinder: Rosch-Buch, Scheßlitz
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier
Printed in Germany

ISBN-10 3-8350-0223-6
ISBN-13 978-3-8350-0223-4

# Foreword

"Knowing, in order to predict". It was this leitmotiv alone that Auguste Comte, in the 19th century, considered suitable for the then rapidly developing empirical sciences. This view remains unchanged until today—not only in the empirical sciences themselves, but also in the philosophy of science. A scientific theory is and should be evaluated primarily on the grounds of its capacity to correctly predict observable phenomena. The aim of predicting takes precedence over the other important aim of science, namely to produce and purposefully manipulate phenomena by technical means in the laboratory, moreover, it even includes the latter. For if scientists successfully produce and manipulate certain phenomena in an experiment, they can ipso facto predict how that experiment will evolve under certain conditions.

We live in a scientifically-dominated world: The more science progresses, the more important correct scientific predictions become. To a sometimes even fatal extent we have made ourselves dependent on science and its results. Our scientifical-technological interventions into nature, yet also into social processes, cover ever larger spatial and temporal distances, and the consequences are ever more drastic given the increasing effort that would be required to reverse the effects—if that is possible at all. That is the reason why we ought to be very well informed about the consequences of our actions, in particular those based on science and technology.

The sciences, under pressure from economic constraints, tend to show themselves at their best, praising their "unlimited" possibilities. Its limits, in particular its forecasting limits, that is something science talks about only reluctantly. The philosophy of science, too, has hardly dealt with this topic so far. This will probably change with the impressive study by Gregor Betz.

This study puts a fundamental insight into action. For, alongside the growth of ordinary knowledge in the sciences, grows the realization that the sciences in fact face rather tight boundaries regarding their predicitive capacities. We have well justified, partially mathematically underpinned empirical knowledge about when, and why, and to what extent sciences fail predictively. This knowledge only has to be deduced from the results of different disciplines, it has to be systematized and, eventually, applied to the analysis of the predictive failure of individual sciences. That is exactly what Gregor Betz does, in an admirably thorough and transparent manner.

Those disciplines which cannot banish the investigated phenomena to the laboratory suffer in particular from forecasting limits. The reader of this book learns why. Last but not least, this affects economics. The failure of economic forecasts is the main subject of Gregor Betz's study.

This is more than just an illustrative example which allows one to discuss the reasons of predictive failure in a particularly transparent way. It is an extremely sensitive example of utmost importance. Public debates about economic policy are obviously dominated by economists who declare with great confidence and hardly contested authority which policy measure will cause which consequence. In short: Economists attempt to predict the macroeconomic development of a national economy, or even the entire world economy. Yet, can macroeconomic developments be properly predicted in the first place? Gregor Betz answers in the negative, warranting his answer with conclusive arguments.

This alone is quite an achievement. And many philosophers of science would be content with such a result. Not so Gregor Betz. For the realization that economics faces insurmountable forecasting limits provokes a highly explosive follow-up question: What are the implications of economic processes being essentially unpredictable for economic policy-making? What does this mean for the rationality of economic policy decisions? Can policy-makers rationally accept the risk which consists in all economic policy measures being blind to the future? And how should citizens, both as the political sovereign in a democracy and as the ones being affected by economic policies, react to the impossibility of economic predictions? This study, instead of avoiding this challenge, tries, in its third part, to confront it with partially surprising, even provocative theses.

An enormously important debate is kicked off. I wish Gregor Betz's book many readers and a — at any rate controversial — discussion.

Holm Tetens

# Preface

Having completed this study almost two years ago, here is its quintessence as I would put it today: This book is about the future. And this book is about us. Our limits; and the difficulties of acting rationally given these limits. Its message carries a critique of a Weltanschauung according to which "we can make it", in brief: hubris. Humility, in contrast, must be the companion of all deliberations about our actions and their long-term consequences.

This said, I would like to place the following notes and acknowledgments in front of the study, touching a more personal string.

"What do you make of economic arguments like those we hear in talk-shows," I asked the student who sat next to me in the bus, "Do you have an idea about how we could verify recommendations regarding economic policy?" Not too sure, either, my neighbor replied: "Statistics, I guess." It was the summer of 1998 and the bus full of undergraduate students was heading for La Villa, Italy, in the heart of the Dolomite Alps where we were going to attend a summer school. "Sustainable development" was the course I had chosen.

The Dolomites are a landscape without horizon. There are but saw-like rows of clear and sharp peaks that attempt to cut the sky in two, waiting to be vanquished. It was at the end of two remarkble weeks that charismatic professor F. J. Radermacher told us to study economics in order not to be fooled by so-called experts—though these, being an economist himself, were not precisely the words he had chosen.

That had, curiously enough, never occured to me before. Back home, the first thing I did was to buy a macroeconomic textbook (Samuelson & Nordhaus, by chance). I devoured it. The mathematical analysis applied to social phenomena, the diagrams, the entire mode of reasoning; never having been tought economics at school I felt as if a door had been pushed open. I was captivated.

Ever since these days my relation to economics has been highly ambivalent: fascination and admiration of its clarity and its rigor on the one hand, doubt and scepticism in the light of its (unacknowledged) limits, sometimes even contempt for its ideological misuse on the other hand. The decision to make this ambivalence one of the central topics of my philosophical studies ripened and was finally reached not long after La Villa when I attended my first lecture on the philosophy of science: a lecture by Holm Tetens at Freie Universität Berlin, entitled "Limits of science".

Holm Tetens was a teacher, and eventually a friend, who shared my interest in understanding both the limits and the role of economics. It is him, first and foremost, whom I thank—for his curiosity, his marked preference for sharpened theses, his healthy scepticism vis-à-vis philosophy, the time spent in discussion with me.

There are other people and institutions to whom this book owes a lot and who have helped me at very different stages of its genesis. Johannes Münster, for instance, urged me to back up my empirical statements much more carefully at an early stage; Julian Reiss gave me reassuring feedback in the final phase. Nancy Cartwright enabled me to spend six months at the Centre for the Philosophy of Natural and Social Sciences in London (a stay generously financed by the German Academic Exchange Service) where we discussed in particular chapter 7 and where I wrote the third part. Collaborating with Stefan Rahmstorf, I visited the Potsdam Institute for Climate Impact Research which was not only important with regard to my climatological case studies but gave me an opportunity to test my claims regarding probabilistic decision frameworks.

Throughout my studies, the Studienstiftung des deutschen Volkes has supported me: not only financially but, even more importantly, ideationally by allowing me to participate in summer schools, language courses and PhD seminars where I faced so many challenging questions and views, met interesting people, made friends. Thank you!

I dedicate this book to my parents.

Gregor Betz

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Economic policy making literally matters to all of us. Economics is supposed to provide the necessary knowledge which enables policy-makers to take effective economic policy decisions. This is why it is so important for all of us to understand how reliable this kind of knowledge is. What can we reasonably expect from economics and where are its limits? One particular scientific objective in the policy-context is to make correct predictions. And the following inquiry is about the limits of such macroeconomic predictions as well as about their sociopolitical consequences.

This study's overall line of reasoning is subdivided into three steps. Its first part consists in assessing the predictive performance of macroeconomics and is going to diagnose that economic forecasts fail consistently (part 1). As a next step, I will present several alternative, non-exclusive and mutually reinforcing explanations for macroeconomic forecast failure (part 2). The diagnosed and explained forecast failure finally has important implications for rational decision making and, in particular, compels us to revise the way we discuss and eventually take economic policy decisions (part 3).

This introduction, in view of preparing the reader for the argumentation which unfolds over the three main parts, first of all exposes and to some extent justifies the *type* of my argumentation. By presenting a very short theory of forecasting, it introduces, in a second place, the terminology and some basic principles which are going to be applied throughout the whole study.

## 1.1 The general approach of this inquiry

In one sentence, this study might be described as a piece of comparative, empirical-analytical theory of science in the tradition of the Enlightenment. The further explication of these three characteristics will clarify my general approach.

*Theory of science as an empirical-analytical science*

Describing my approach as empirical-analytical entails that it is not a priori: it is not classical epistemology. For I believe that there are no relevant a priori limits neither of economics nor of any other empirical science. In doing so, I take a statement (about a limit of science) to be a priori, if and only if it holds irrespective of what the world is like, i.e. if it can be derived from purely analytical statements only.

As a first point to notice, a priori arguments that establish fundamental limits of science typically fail. As early as in the late 1920s, after the first heydays of economic forecasting, Oskar Morgenstern argued in *Wirtschaftsprognose: Eine Untersuchung Ihrer Voraussetzungen und Möglichkeiten* for a very strong conclusion, namely that "economic prediction is principally impossible for factual reasons" (Morgenstern, 1928, p. 108)[1]. He thus not only attempted to show that there are a priori limits of economic forecasting but that economic forecasting is altogether impossible for a priori reasons. In other words, Morgenstern argued that economic forecasts necessarily fail. Although none of his arguments actually warrant this strong conclusion (see Betz, 2004) some of them can be reinterpreted as explanations of forecast failure and will be referred to in part 2.

More recently, Nicholas Rescher presented in *Predicting the Future* some a priori arguments which seem to prove some fundamental limits of prediction. The following paradox is meant to show that the very idea of a perfectly predicting machine, i.e. a machine that answers *any* question about the future correctly, is inconsistent,

> [We] could wire our supposedly perfect predictor up to a bomb in such a way that it would be blown to smithereens when next it answers 'yes'. And we now ask it: 'Will you continue active and functioning after giving your next answer?' (Rescher, 1998, p. 213)

Additionally, Rescher (1998, p. 218) presents a classical infinite regress of justification in order to demonstrate that even if there were a perfect predictor, we would never come to know it (or her or him). Here is a concise reconstruction of that argument,

(1) We only know that *P* is a perfect predictor, if it is justified to belief that all of *P*'s predictions will turn out true.

(2) Such a warrant can only be given by a perfect predictor who is known as such.

[1] This as well as all the following quotes with German references are my translation. The original version of longer quotes is given in the appendix.

(3) *Thus:* We cannot know whether something or somebody is a perfect predictor or not.

I will skip the discussion of whether these arguments are sound and really highlight a priori limits of prediction[2]; the point I would like to stress is rather: their conclusions do not tell us anything about the possibility of economic prediction in particular. They are of no relevance with regard to our central questions since what we aim at in economic forecasting is obviously not the construction of a perfect predictor.

In fact, one can give several arguments which conclude that there are no relevant a priori limits of forecasting and which thereby show that arguments like those of Morgenstern and Rescher have to fail. Consider, first of all, the argument from abstractness: It states that if there were any a priori limits of forecasting, then they would have to be so abstract in order not to be empirical that they would not be interesting or relevant. For any such a priori statement would by definition of "a priori" apply to our world as well as to every other possible world, as for instance the possible world consisting of merely one planet completely covered by moss. Whatever the forecasting limits are that our and that green world have in common: they are of no interest when we want to assess and understand economic forecasting limits.

A second argument is based on the following idea: One can consistently describe and even simulate a scenario where economic agents behave exactly like presupposed in a simple economic model, say for instance Solow's growth-theory[3]. In that scenario, economic prediction is feasible, namely by making use of Solow's theory. In addition, all analytically true statements hold (by definition) in that scenario, too. But then, the impossibility of macroeconomic forecasting cannot be derived from purely analytical statements only. The scenario just described is a counter-example. The complete argument can be reconstructed in a model-theoretic way, but readers not familiar with model-theory may just skip the reconstruction.

(1) There is an artificial, i.e. computer-simulated, world where

- households and firms behave as Solow's growth-theory presupposes, and, thus, Solow's growth-theory is true,

[2] Just notice that the paradox only shows that there cannot be a vulnerable (!) perfect predictor. What if the predictor were indestructible? As to the second argument, premiss (2) is anything but obviously true for why should one not know by the very functioning of the predictor (suppose it is a machine) that it makes correct predictions — maybe not in our complex, but at least in a somehow simpler and more ordered possible world.

[3] Which may count as the first modern theory of economic growth and is exposed in any textbook about growth theories.

- macroeconomic prediction is feasible, i.e. it is not true that macroeconomic prediction is impossible ($\neg p$).

(2) This artificial world is a model (in a model-theoretical sense) of our economic vocabulary and can be expanded to a model, $\mathcal{M}_s$, of our ordinary language that makes Solow's growth-theory as well as all analytical statements, $A$, true.
(3) *Thus:* Not for all models $\mathcal{M}$: $(\models_{\mathcal{M}} A) \Rightarrow (\models_{\mathcal{M}} p)$, i.e. $A \not\models p$.
(4) $A \vdash p \Rightarrow A \models p$. (Consistency)
(5) *Thus:* $A \nvdash p$.

Quine's attack of the analytic-synthetic distinctions yields, last but not least, a third argument against a priori limits of forecasting. Quine argued — and most philosophers would say: successfully — that there is no sharp distinction between empirical and analytical statements. Our net of beliefs faces empirical tests as a whole and *any* of its components could principally be revised in order to re-establish coherence. Clearly, if there are no statements which are true by virtue of their meaning and irrespective of what the world is like, then statements about limits of science, in particular, cannot be of that kind, either.

(1) There are no analytical statements which cannot be revised given new empirical evidence.
(2) A priori limits of science which hold irrespective of empirical facts are by definition analytical statements which cannot be revised given new empirical evidence.
(3) *Thus:* There are no a priori limits of science which hold irrespective of empirical facts.

Besides of being non-a priori, my approach is also non-metaphysical. What I have in mind here is not only the claim that there are no metaphysical limits of forecasting but also that one cannot explain forecast failure in metaphysical terms. Successful explanations of forecast failure are empirical.

Consider for instance the metaphysical scenario Nancy Cartwright depicts in *The Dappled World* according to which the macroscopic world "is for the most part an unruly mess". (Cartwright, 1999, p. 110) In the Dappled World, predictability is the exception — rather than the rule. But might this general statement serve as an explanation of forecast failure?[4] I do not think so, mainly for two reasons. First of all, the metaphysical scenario does not tell us why the planetary system on the one hand is largely predictable whereas the economic system, on the other hand, is not. Why is that the case and not the opposite? Why is the

[4] I do not want to suggest that Cartwright actually considers it as an explanation.

economic system not an exception to the metaphysical rule, too? What makes the sun-system an exception? Since the metaphysical scenario does not answer these questions, it is hardly a satisfying explanation of macroeconomic forecast failure.

In addition to these difficulties, I cannot see how somebody could come to know that the World is such-and-such, for instance an unruly mess. As long as we interpret Cartwright's scenario as genuinely metaphysical, we have no reason to believe it. Alternatively, it could be read as a general statement which summarizes the typical experiences that humans engaged in all kinds of epistemic practices have made for centuries. But as such, it would of course be an empirical statement which solely rephrases instead of explains that forecasts typically fail.

Being non-a priori and non-metaphysical, this study is written from a, as Putnam (1981) called it, "internalist perspective". All its results, including the diagnosed forecast failure and its explanations, hold conditional to the way we conceive the world in general and the economic system in particular. In other words: It might be that one day an ingenious scholar comes up with a revolutionary new way to talk about economic phenomena and that what is referred to as "economy" according to that new language exhibits stable regularities which enable us to make correct predictions. Possibly none of the reasonings presented in part 2 would be true within that new framework. On the other hand, our results, in particular the obstacles to economic forecasting identified in part 2, are real and urgent as long as we talk about the economy the way we talk about it today and have talked about it at least during the last 200 years. And, of course, as long as future economies are sufficiently similar to ours.

Equally, the notions "truth", "world", "reality" and so on are used in their internalist, non-metaphysical meaning. In line with my philosophical conviction that metaphysical notions (of truth etc.) are as relevant to our lives as the skeptical hypothesis (that *everything* might just be dreamed etc.) — namely of no relevance at all —, this study is also a demonstration of the ultimate importance of the internalist perspective.

A further characteristic of my approach which the term "empirical-analytical" tries to capture is that case studies will play an indispensable role in the course of the argumentation. However, this study is not just a piece of empirical science, not just another empirical study of economic forecast performance. For the argumentation, notably in parts 2 and 3, seriously involves methodological reasonings, too: The explanatory arguments are typically based on empirical *and* methodological premisses. As I will make explicit in due course, some of these arguments rely on Cartwright's analysis of the scientific method, specifically her work on the role of models as mediators between experiments and theories and the concept of nomological machines.

*Comparative theory of science*

As I have just said, case studies will form an important ingredient of this inquiry. These cases will not only be drawn from economics but from a different science, too, in order to enable us to make interdisciplinary comparisons. Whereas comparing forecast limits of economics with forecast limits of another science will be illuminating, comparing *explanations of* forecast limits will be indispensable. Since examining whether an argument that seems to explain economic forecast failure applies *mutatis mutandis* to another science with similar or different forecasting limits is a way to check whether the argument really explains or not, explanations should be cross-disciplinarily stable. If, to cite an example, poor data quality explains forecast failure it should do so in economics as well as in physics: i.e. it should not be the case that forecasts in some branch of physics perform relatively well irrespective of the underlying data's quality.

In order to secure some coherence, we will pick the non-economic case studies from one (admittedly large) discipline only: thermodynamics. This choice combines the advantages of coherency and manifoldness as cases ranging from Lavoisier's experiments to modern climate modeling fall under that science. The analogies we will be able to draw between thermodynamic and economic case studies will prove to be fruitful.

*Theory of science in the tradition of Enlightenment*

This characterization is about the fundamental objectives of my study and consequently its addressees. Although, as we have seen above, the argumentation involves methodological reasonings, the ultimate aim is not a methodological one. That is, I do not try to give a complete account of the methods of economic forecasting nor do I want to reconstruct them in a rational way. By no means do I suggest measures to improve macroeconomic forecasts. This is, I believe, the forecasters' business. Yet, I would of course be delighted to see that one idea or another from this study turned out to help a forecaster to improve his predictions.[5]

Instead, this study forms a part of the philosophical project that aims at assessing critically and in a fair way the limits of science so as to enable interested laymen to make an informed judgment about what one can and cannot reasonably expect from science. It is hence an attempt to increase public understanding of science. Over and above that, this project questions the role science can and should play in our societies, and makes suggestions how to deal with it. Ultimately, it

[5] The first thing to do in order to use this study for forecast improvement would be to assess empirically which of the alternative explanations of forecast error is actually the most important one. As a next step, the forecast methods would have to be modified in order to avoid the problems highlighted by that explanation.

thus arises out of very practical, political concerns and aims at changing — not science, but its place in our society.

In this sense I consider my approach as standing in the tradition of Enlightenment. It is, in the words of Bernard Williams, the Enlightenment's "commitment to honesty and transparency and its rejection of power that falsely presents itself as cognitive authority" (Williams, 2002, p. 231) I subscribe to.

## 1.2 A very short theory of forecasting

*Typology of predictions*

A prediction is a statement on the future. Thus, predictions can be classified according to the type of statement they assert, namely whether it is an ontological, a qualitative or a quantitative one. Correct *ontological predictions* rank among the most famous successes of science: deviations of Uranus' observed orbit from what it was calculated to be according to Newton's laws, for instance, lead to the prediction of a further planet, Neptune, whose eventual discovery counted as a triumph of Newton's theory. To be precise, an ontological prediction is not a prediction of (the emergence of) a new entity. The ontological statement, e.g. Neptune's existence, is supposed to be timeless. What is predicted is, in contrast, the discovery of that very entity. Although ontological predictions are of great importance in natural science, they do not seem to play an important role in economics and are therefore of no particular interest to our study.[6]

*Qualitative predictions* give an account of a future state in classificatory or comparative terms. An individual is predicted to fall under a certain class, "By spring, this company will go bankrupt", or compared with another individual, "Next summer will be drier than this year's". As opposed to qualitative ones, *quantitative predictions* describe the future by making assertions about certain variables. "By spring, share prices of that company will drop by $n\%$" or "Next summer, rainfall will amount to $n$ liter per square-meter" are for example two quantitative predictions — with share prices and precipitation as *predicted variables* respectively. Quantitative predictions, or *forecasts*, as we will use the term, are the kind of predictions our inquiry is concerned with.

A (future) state can be characterized by a variable in different ways which gives rise to different types of forecasts. A *point forecast* predicts the exact future value of the predicted variable, $X$, at time t,

$$_{t-1}X_t^F = x,$$

[6] The prediction of the emergence of a new kind of firm or an entire economic sector might be considered as an economic ontological prediction.

where $t-1$ denotes the time when the forecast is made. An *interval forecast* merely specifies the upper and lower boundaries of the predicted variable's future value,

$$x_1 < {}_{t-1}X_t^F < x_2.$$

Yet, a *directional* (or *trend*) *forecast* specifies the future direction of change of the predicted variable,

$${}_{t-1}X_t^F - X_{t-1} \lessgtr 0,$$

and, finally, a *density forecast* predicts the probability density function the predicted variable will obey,

$${}_{t-1}\mathrm{P}_t^F(X) = \ldots.$$

Such a density forecast, by integrating ${}_{t-1}\mathrm{P}_t^F()$, allows to infer the probability that some interval forecast is correct. An interval forecast whose probability of being correct is known will also be referred to as *probability forecast* (PF).

The above forecasts are *positive forecasts* as opposed to their *negative* counterparts which we obtain by simply negating the positive forecast. Accordingly, a negative point forecast consists in stating that the predicted variable will be unequal to some specific value, a negative density forecasts predicts that a given density function does not characterize the predicted (random) variable.

The forecasts we have considered so far are *categorical*: They state what is going to be the case — full stop. *Conditional forecasts* (CF), in contrast, state what is going to be the case — if some antecedent condition is fulfilled. Let $F$ stand for some categorical forecast and $C$ be some conditional statement, then the conditional forecast $F_C$ is defined as,

$$F_C \iff (C \rightarrow F). \tag{1.1}$$

This definition applies no matter whether $F$ is a point, interval, directional, density or probability forecast and irrespective of what kind of statement $C$ is. Yet, $C$ typically is a statement on the future, too. Another more specific point is that a conditional density forecast entails — due to the logical connection between density and probability forecasts — a conditional probability forecast (CPF), i.e. a forecast of the conditional probabilities of some (future) events.

From the definition of conditional forecasts (1.1) follows that a conditional forecast plus its conditional statement imply a categorical forecast,

$$(F_C \wedge C) \Rightarrow F.$$

If $C$ is a statement on the future, then we say that the categorical forecast is entailed by the conditional one and the *co-prediction* of its antecedent condition.

A co-prediction, generally, is a prediction associated with a categorical forecast. Thus, a statement on the future which is entailed by a categorical forecast given some background knowledge (a model, for instance) is a co-prediction, too.

Forecasts can be classified according to several further characteristics. Consider a forecast, ${}_{t_1}X^F_{t_2}$, which has been made at time $t_1$ and predicts the variable's value at $t_2$. Its *forecast horizon* is the difference $t_2 - t_1$. The time period into which $t_2$ falls is called the *forecast period.* The *forecaster* is the individual or institution who has derived the forecast.

Forecasts are inferred from some past data: the *predicting data.* The *predictive inference* which enables one to derive the forecast from that data is typically based on a *predictive model.*[7] Forecasts can hence be classified according to the different types of predictive models — or generally: methods — they rely on. If a predictive model and some predicting data is not used to infer a forecast but a statement on the past (relative to the predicting data), then this statement is called a *retrodiction.* The performance of retrodictions may justify a predictive model's reliability as long as the predicting data does not already include the fact that ought to be retrodicted. Retrodictions, in order not to be circular and in order to warrant a predictive model's reliability, thus have to be *out-of-sample. In-sample* retrodictions, in contrast, are mere pseudo predictions.

### *Correctness and Credibility*

"Prediction", Rescher (1998, p. 65) puts it, "is always to some extent a leap into the future." Whether or not a predicted statement is true will only turn out (if at all) once it is no longer a prediction. Thus, if the truth-values of predictions can only be evaluated ex post, how do we evaluate forecasts ex ante then? When is it sensible to incorporate a categorical forecast into our body of knowledge, when is it reasonable to base our actions on it, in other words: When should we consider it as *credible* and not merely as a *prophecy*?

To answer these questions, we have to restart with truth. Strictly spoken, a forecast, at least a point forecast, will almost never be true. All one can reasonably expect is approximate[8] truth. A forecast which is approximately true is a *correct* one. A forecast has *failed* if it turned out to be incorrect. Forecasts of a certain kind (defined by the predicted variable, forecast period, similar forecast horizon, method, or ... ) are said to be *successful* if they are typically[9] correct. Forecasts of some kind, accordingly, perform well if they are successful and poorly if not.

[7] That is why Makridakis (1984) says that forecasting consists of "fitting a model to a set of data" (p. 2).

[8] What this exactly means is one subject of the next subsection.

[9] Again, to be specified below.

So far, so good, but how do we get from this to forecast credibility? I suggest to assume the two following principles: First of all, the consistent failure of some kind of forecasts is a sufficient condition for not considering such forecasts as credible. And secondly, their consistent success is sufficient for their credibility. Accepting or rejecting a forecast as credible therefore involves an inductive step which bears the risk of error. But here, a

> justification of induction need not issue in any *preassurance* of its success. Initially, it is a matter of *faute de mieux* argumentation [...]. And *ultimately* we have recourse to retrojustification [...]. (Rescher, 1998, p. 66)

At the end of the day, the justification of the two principles above is a pragmatic one.

Now, what if forecasts of some kind have neither consistently failed nor performed well? What if there is no empirical evidence at all? In such cases, further evidence becomes crucial: A forecast, then, is as credible as it is sensible to accept the assumptions, i.e. the data and the predictive model, it is derived from.[10]

As to conditional forecasts, the first thing to note is that a CF is correct whenever its corresponding categorical forecast is correct, for (1.1) implies that $F \Rightarrow F_C$. So conditional forecasts seem to become true more easily than their categorical counterparts. Moreover, every categorical forecast can be transformed into a conditional one which is necessarily correct. Just take a set of assumptions which are sufficient to deduce the categorical forecast as the antecedent and there you go. But in the sense that such a conditional forecast is an analytical statement, it does not tell us anything about the world.

The relatively weak truth-conditions of CFs are the reason why the earlier definition of credibility cannot simply be extended from the categorical to the conditional case. To see why, consider a conditional forecast whose antecedent condition has not been realized yet; for example the proverb of the German city of Marburg which says that if one of the city's graduated theology students is a virgin, the leaning tower of one of the city's churches will sit up. Such a conditional forecast has on any occasion been correct and hence successful because the antecedent condition has never been met. But should we therefore incorporate it into our background knowledge, i.e. consider it as credible? Everybody agrees: That would be premature. Then, what does it take a conditional forecast to be credible?

As long as the antecedent condition is not satisfied, success is not sufficient to consider a CF as credible. Now, what if the antecedent condition is satisfied? First

[10] In this respect Morgenstern was wrong when stating that "there is no other benchmark for the prediction than its truth" (Morgenstern, 1928, p. 115).

of all, this might give rise to a falsification of the conditional forecast. From (1.1) directly follows that

$$(\neg F \wedge C) \Rightarrow \neg F_C,$$

that is the conditional forecast is incorrect if its antecedent condition is met but the categorical forecast nevertheless fails. If this has consistently been the case for similar CFs, then we should consider the conditional forecast as non-credible. On the other hand, if the corresponding categorical forecasts have consistently turned out to be correct as long as the antecedent conditions were fulfilled, then, I suggest, we have good reason to consider the CF as credible. For what we ultimately rely on when we accept a CF is that once its antecedent conditions are fulfilled, the resulting categorical forecast will be correct. That is what CFs are supposed to tell us. The CF is no more and no less credible than the corresponding categorical forecast in case the antecedent condition is satisfied; those are the test-cases of every CF.

Similar to the unconditional case a CF might be considered credible for other reasons than those just cited. If one has very good reasons to accept every assumption that was necessary to derive the CF, then one has, of course, equally good reasons to accept the CF itself.

These reasonings have important implications. In some sense, the availability of credible categorical forecasts has turned out to be a precondition for the availability of credible conditional forecasts. For assume that there were no credible categorical forecasts at all, i.e. forecasts consistently failed irrespective of the circumstances. This would entail that whenever the antecedent conditions of a conditional forecast had been satisfied, the corresponding categorical forecast would have failed, too. And as a consequence, conditional forecasts would either consistently fail and not be considered as credible or not be tested at all so that their credibility would purely rest on the credibility of their underlying assumptions.

Eventually, I shall make some remarks with regard to negative forecasts. As every (informative) positive forecast entails at least one negative forecast, the availability of credible positive forecasts implies the availability of credible negative forecasts. Or, the other way around, no credible negative forecasts without credible positive ones. Still, the inverse does not generally hold. One might not be able to derive a credible positive forecast in spite of having excluded some possible future states. The failure of negative forecasts thus is a much stronger claim than the non-availability of credible positive forecasts.

*Absolute error measures*

We have previously defined correctness of a forecast as approximate truth and success as consistent correctness. This subsection is supposed to make these concepts empirically operational by providing a measure of forecast performance.

To assess a positive categorical point forecast ex post all one seems to have to do is to calculate the difference between the predicted value and the actual outcome, i.e. the individual error of the forecast, $e_t = {}_{t-1}X_t^F - X_t$. The smaller the error, the closer the forecast comes up to the truth. With a threshold which defines how close is close enough for approximate truth, we thus can evaluate forecasts empirically.

Yet, things are slightly more complicated as a good or a poor performance of a single forecast might simply be due to chance. We have to know that some kind of forecast is consistently correct in order to infer its credibility. Accordingly, we may not only stick to one forecast, but have to evaluate a whole sequence of forecasts, such as for example $X^F = \langle {}_{t-1}X_t^F, {}_tX_{t+1}^F, \ldots, {}_{t+n-1}X_{t+n}^F \rangle$.[11] What is the error-term which characterizes not the individual forecasts' but their overall performance?

A first way to think about this problem, exploiting the spatial connotation of approximate truth, is in geometrical terms: What is the distance between the forecast sequence $X^F$ and the corresponding sequence of actual outcomes? Well, that is the length, according to some norm $\|\ldots\|$ defined on the $n$-dimensional vector space, of their difference vector, which, in turn, is nothing but the sequence of individual forecast errors, $e = \langle e_t, e_{t+1}, \ldots, e_{t+n} \rangle$. Thus, by defining a norm for the error sequences one automatically creates a metric which measures the distance between some forecast sequence and the corresponding actual outcomes (explicitly: $d(X^F, X) := \|e\|$) and which, consequently, quantifies the overall forecast performance.

Every metric and every norm by definition fulfills a fundamental equation, the so-called triangle inequality. Now, the geometrical stance towards forecast assessment, i.e. the measurement of forecast performance via the error sequence length, assumes that it is reasonable to apply the same inequality to forecast measures. Is that the case? Applied to forecast evaluation the triangle inequality reads,

$$\|e\| + \|e'\| \geq \|e + e'\|,$$

where $e$ and $e'$ are the error sequences corresponding to two different forecast sequences $X^F$ and $X^{F'}$. The inequality then states that the overall error of the

[11] Whenever I loosely talk about the performance of this and this forecast in the following chapters, it is ultimately a sequence of forecasts (determined by the context) I refer to.

combined error sequence[12] cannot be greater than the sum of the two sequences' overall errors. This excludes in particular that large errors are over-proportionally 'punished': For assume that the individual errors of the combined sequence are twice as large as the uncombined forecasts' errors. Then the triangle inequality prohibits to assign a more than twice as large overall error to the combined sequence than to each of the uncombined sequences. It is anything but clear why this should be the case and some forecast error measures do indeed violate the triangle inequality, as we will see below. Although the triangle inequality is clearly no mantra for forecast evaluation, the geometrical stance still provides an intelligible interpretation of some forecast error measures and a helpful framework to organize them.

The natural way to measure distances and length is by adopting a Euclidean metric. Applied to forecast performance, this yields the root square error measure (RSE),

$$d(F, X) = \|e\| = \text{RSE} := \sqrt{\sum (e_i)^2}.$$

However, a further obstacle arises immediately: The forecast sequences we want to compare are rarely of the same length whereas that is what is required by the metric-approach. It is not possible to define a metric for vectors of different spaces, but we can try to neutralize the effect of the dimension on the overall measure. Specifically, we demand that an adjusted metric attributes the same overall error to some error sequence $\langle e_t, e_{t+1}, \ldots, e_{t+n}\rangle$ as to the doubled sequence $\langle e_t, e_{t+1}, \ldots, e_{t+n}, e_t, e_{t+1}, \ldots, e_{t+n}\rangle$. The root mean square error (RMSE) modifies the RSE in order to satisfy this condition,

$$\text{RMSE} := \sqrt{\frac{\sum (e_i)^2}{n}}.$$

Since $\lambda\Delta$ is a metric if $\Delta$ is a metric, RMSE defines a metric (for each $n$-dimensional space) in the mathematical sense. Although the formula to calculate RMSE is rather complicated, we have shown that its construction within the geometrical framework is fairly straightforward. Another straightforward way to define a metric is to sum the individual absolute errors and to account for size by dividing by $n$. This yields the so-called mean absolute error,

$$\text{MAE} := \frac{1}{n} \sum |e_i|.$$

Like RMSE, MAE is adjusted for the forecast sequence's dimension. MAE and RMSE are probably the two most frequently used error measures in the literature

[12] Which must not be confused with the error sequence of the (combined) mean forecasts.

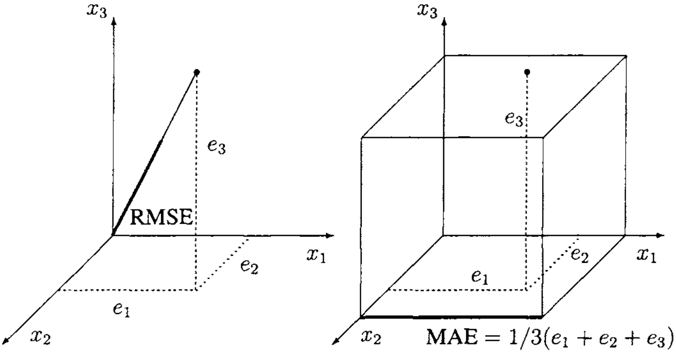



Figure 1.1: Geometrical illustration of RMSE and MAE in the three-dimensional space of error sequences $\langle e_1, e_2, e_3 \rangle$.

on economic forecast evaluations, for sure in part 1 of this study. Due to the squaring of the individual errors, RMSE 'punishes' large errors more than smaller ones.[13] Consider for example the two error sequences $\langle 1, 1, 1, 1 \rangle$ and $\langle 2, 2, 0, 0 \rangle$. Obviously, they both have the same MAE. But they differ in their RMSE. The second sequence with fewer but higher error terms has a RMSE of $\sqrt{2}/2$ while the first's RMSE of $1/2$ is clearly smaller. As a rule of thumb, $\mathrm{RMSE} \approx 1.25 \cdot \mathrm{MAE}$ (Öller & Barot, 2000, p. 30). Figure 1.1 illustrates RMSE and MAE by plotting their geometrical interpretation for a three-dimensional error sequence.

We could go on inventing and discussing different metrics for the next hundred pages. Yet, the metrics above are the main absolute forecast error measures which fall under the geometrical perspective. That is the reason why we will now change our point of view: The second, alternative stance towards forecast assessment is the statistical one. According to this perspective, a forecast sequence's performance is still determined by the corresponding error sequence $\langle e_t, e_{t+1}, \ldots, e_{t+n} \rangle$. The error sequence is, however, no longer considered as a vector in an $n$-dimensional space but as a finite sample of some random variable $e$, i.e. the forecast error, instead.[14]

This new point of view allows us to characterize a forecast error statistically, namely by its expected value (the mean error), its variance and its standard deviation. The mean error contains information about whether the predicted variable is systematically over- or underestimated. A forecast with a non-zero mean error

13 Yet not too much more in order to comply with the triangle inequality.

14 Note that the inductive step discussed in the previous section is to some extent built into this approach: Forecasts of some kind are supposed to be characterized by one single random error variable. Now assume that "forecasts of the same kind" is defined without reference to the forecast period. Hence, the empirical investigation into the random error variable, i.e. the statistical evaluation of a finite sample, yields results that *eo ipso* characterize the error of future forecasts.

is called *biased*. We may consider the mean error as a measure for the forecast's overall *accuracy* Hendry (2001). This so defined measure of accuracy does not tell us anything about the forecast's *precision*. Consider for example two forecasts with error samples $\langle 1, 0, -1, 0\rangle$ and $\langle -1, 4, -5, 2\rangle$. They both are unbiased and are consequently perfectly accurate. But the first forecast is more precise for it is less volatile than the second one. The natural statistical measure for this volatility is the variance of a random variable: It is defined as the expected deviation from its mean. The variance as a measure for volatility might also be replaced by its square root, the so-called standard deviation. Let $e$ be the random error variable of a forecast, then

1. the forecast's accuracy is measured by the mean error $\mathrm{E}(e) = \mu = \frac{1}{n}\sum e_i$
2. the forecast's precision is measured by the variance of its error $\mathrm{V}(e) = \mathrm{E}(e-\mu) = \sigma^2$ or by its standard deviation $\sigma$.

Following Hendry (2001), we can join these measures for accuracy and precision together in "a natural combination" (p. 27) in order to obtain an overall measure for the forecast's performance. This combination is,

$$\phi = \sigma^2 + \mu^2,$$

where $\phi$ denotes the new error measure. But what is $\phi$? How does it relate to the other forecast error measures? Is it a metric? To answer the first question, we substitute the formulas for standard deviation and mean into the above equation and obtain,

$$\phi = \frac{1}{n}\sum(e_i)^2,$$

which is nothing but the mean square error (MSE), the square of RMSE! But is MSE a metric? No, because MSE does not satisfy the triangle inequality as an easy counter-example shows: Let $e = e' = \langle 2\rangle$ be the error sequences of two forecasts $X^F$ and $X^{F'}$, then

$$\mathrm{MSE}(X^F) + \mathrm{MSE}(X^{F'}) = 8 < 16 = \mathrm{MSE}(X^F + X^{F'}).$$

The fact that MSE is no metric is no disadvantage as such as we have said earlier; it merely indicates that MSE does not fall under the geometrical stance. But can our metrics on the other hand be integrated into the statistical approach? That is at least the case for the RMSE because the MSE is simply its square. Accordingly,

$$\mathrm{RMSE} = \sqrt{\mu^2 + \sigma^2}.$$

If the forecast is unbiased, that is $\mu = 0$, then the RMSE is equal to the forecast error's standard deviation $\sigma$. But in contrast to the RMSE, the MAE cannot be expressed as a simple function of the forecast error's mean and variance.

So far, we have become familiar with the most important forecast error measures. But the statistical stance does more than simply provide error measures. It helps us to clarify at least partly the problem of the inductive inference which is involved when judging whether some forecast is credible or not. That inference from consistent success to future success and thus current credibility is invalid if, for instance, consistent (past) success is merely due to chance. The question is, in terms of the statistical approach: Is the past success of this-and-this kind of forecasts *significant*? What we need is a measure of how representative our sample really is with regard to a certain conclusion — the so-called null hypothesis $H_0$ — we inferred. The statistical significance is such a measure: Making more or less restrictive assumptions concerning the examined random variable, a positive significance test establishes for a given sample with a predefined probability (in most cases more than 95%) that $H_0$ has not been inferred spuriously, i.e. is not in reality false but merely passed the statistical test because of pure chance or random fluctuation of the examined variable.

Such significance tests that can be applied to forecast evaluation have recently been developed during the 1990s (Mariano, 2002). They differ in making different assumptions concerning the statistical properties of the forecast error. And because these significance tests are rather new, not all the reported studies in part 1 have implemented such tests.

### *Relative error measures*

The error measures we have discussed so far characterize absolute errors. They do not take into account the size of the predicted variable, the forecast period, its volatility, nor any further characteristics. This is, in general, sensible because it allows to study how forecast errors vary when for instance the predicted variable's volatility increases. Yet, we are in a position where we want to compare the performance of forecasts which have been made in different periods and countries, and whose predicted variables thus are of different size and volatility. The aim of our comparisons is not to study the effect of volatility on forecast errors but to see whether forecasts have improved over time or whether one forecaster has consistently outperformed others. Ideally, the effects of all relevant factors which possibly distort the result would be taken into account by the statistical analysis itself — but the available sample size is not large enough to do so. That is why forecasters and econometricians try to adjust the forecast error measures to the different possibly relevant factors. I will now briefly outline how error measures can be adjusted to variable size and volatility.

The standard adjustment to the predicted variable's size consists in expressing individual forecast errors in relative terms (percentage) when calculating the overall error. A corresponding modification of the MAE, the mean absolute percentage error, does so,

$$\text{MAPE} := \frac{1}{n} \sum |\frac{e_i}{X_i}|.$$

An error measure which divides the absolute errors by the predicted variable's average value, $\overline{X}$, instead of its corresponding current value is another way to adjust to predicted variable size,

$$\text{MAPE2} := \frac{1}{n} \sum |\frac{e_i}{\overline{X}}|.$$

Both error measures are metrics in the strict mathematical sense. Likewise modifications can also be applied to the RMSE and the MSE.

There are two ways to adjust an error measure to the predicted variable's volatility. Both adjustments assume that the more volatile the predicted variable, the higher the difficulty to forecast it. The first method consists in dividing the error measure, say RMSE, by a measure for the predicted variable's volatility, most commonly its standard deviation. The obtained fraction weights the RMSE according to the "difficulty to predict" the variable. The second method compares the forecast under examination with some naïve benchmark forecast: a "no change" or an "average change" extrapolation, for instance. In this second case, the RMSE of the evaluated forecast is not divided by the variable's standard deviation but by the RMSE of the naïve benchmark. The lower this fraction, the better the forecast performed compared to the naïve model. A fraction greater than one indicates that the forecast is outperformed by the naïve benchmark. Now, in the end, both methods amount to the same because the naïve model will generally perform much better in calm than in turbulent times: Division by the naïve model's error is hence another way to take the difficulty to predict a certain variable into account. But naïve models are not only a useful device to account for forecasting difficulty, they also yield a natural benchmark any credible prediction should beat.

The basic error measures introduced hitherto plus the different possible combinations of adjustments together yield several dozens of forecast error measures. Yet, these are by no means all error measures which can be used to assess forecast performance.[15] In fact, it makes perfect sense for forecasters and the users of forecasts to assess forecast errors by very specific measures for such errors might

[15] A further prominent family of more elaborated measures for example consists of variations of Theil's U-measure, see Loungani (2001, p. 422).

cause very different, case-specific costs: Error measures can also be thought of as cost functions (Granger, 2001). But, over and above that, empirical evaluations of macroeconomic forecasts typically make use of the basic measures which I have introduced, namely RMSE, MSE and MAE, and their adjustments.[16]

[16] Whereas econometricians underline the advantages and make use of the MSE (see the contributions to Hendry & Ericsson (2001) and Clements & Hendry (2002*a*), in particular McCracken & West (2002)), most empirical studies, as for example those reported in the *International Journal of Forecasting*, make use of RMSE and MAE. Camba-Mendez et al. (2002, p. 391) confirm this observation.

# Part I

# Assessing the predictive limits of economics

# Chapter 2

# Recent performance of economic forecasts


## Summary

This chapter assesses the current performance of macroeconomic forecasts. It reviews the findings of empirical forecast evaluations and discusses them thoroughly while distinguishing absolute and relative performance, the performance of specific methods and that of different forecasting institutions. All things considered, the evaluation diagnoses a comprehensive failure of macroeconomic forecasts. The last section summarizes the main conclusions that can be drawn from the discussed studies.


## 2.1 General performance of macroeconomic forecasts

The assessment of economic forecast performance in this and the following chapter will focus on categorical point-forecasts of main macroeconomic variables, namely GDP and price level, as well as on some financial forecasts. Macroeconomic forecasts typically predict growth rates instead of absolute values: nominal GDP growth (or simply "growth"), price level changes (the "implicit price deflator" or simply "inflation") and real GDP growth which, being defined as nominal growth minus inflation, links both of the former. Unless explicitly specified, the forecast horizon of all forecasts in this and the following chapters is equal to one year. This section reports the forecasting literature's findings on macroeconomic forecasts' general performance before the subsequent sections discriminate between different forecasters, and predictive methods. Table 2.1 gives an overview of the empirical studies the assessment is based on.

*Forecast performance in absolute terms*

The forecasting literature confirms, but not without exceptions, the following general statements concerning the absolute errors of macroeconomic forecasts: The

| | Evaluated Forecast | | | Evaluation | | |
|---|---|---|---|---|---|---|
| | GDP/CPI | Period | Country | prim. | sec. | Reference |
| 1 | +/+ | 1960-75 | US, UK | – | + | Armstrong (1984) |
| 2 | +/+ | 1953-90 | US | + | + | Zarnowitz (1992) |
| 3 | +/+ | 1968-90 | US | + | + | Zarnowitz & Braun (1992) |
| 4 | +/+ | 1971-91 | US | + | + | McNees (1992) |
| 5 | +/+ | 1965-89 | US | + | – | Joutz & Stekler (2000) |
| 6 | +/+ | 1971-98 | 13 | + | – | Öller & Barot (2000) |
| 7 | +/+ | – | – | + | + | Makridakis & Hibon (2000)* |
| 8 | +/+ | 1991-2000 | 5 | + | – | Blix et al. (2001) |
| 9 | +/+ | 1989-98 | 63 | + | – | Loungani (2001) |
| 10 | +/+ | 1970-2000 | US, UK | – | + | Burns (2001) |
| 11 | +/– | 1970-2000 | UK | + | + | Osborn et al. (2001) |
| 12 | +/+ | 1990-2000 | UK | – | + | Coyle (2001) |
| 13 | +/+ | – | – | – | + | Fildes & Ord (2002)* |
| 14 | +/– | 1990-98 | 4 | + | – | Camba-Mendez et al. (2002) |
| 15 | +/– | 1966-2001 | G | + | – | Dicke & Glismann (2002*b*) |
| 16 | +/– | 1964-2001 | G | + | – | Dicke & Glismann (2002*a*) |
| | 16/12 | 1950-2000 | | 12 | 9 | Total: 16 |

Table 2.1: Literature basis. Explanation: Other countries than those explicitly mentioned include mainly OECD-countries except 9. A "primary evaluation" is a statistical evaluation of a forecast's performance while "secondary evaluation" denotes a survey of such primary evaluations. Studies marked with a "*" discuss so-called forecasting-competitions.

MAE of growth forecasts is slightly larger than 1% while the inflation forecasts' MAE is slightly less than 1%. The corresponding RMSEs are larger, satisfying roughly the rule of thumb mentioned in the introduction (see p. 14 above). If we assume that the long term growth equals about 2%, the MAE is roughly equal to half the size of the forecasted variable.

Let us have a closer look at the studies which establish these conclusions. The initiator of the ASA-NBER survey of US-American forecasters has evaluated the recorded statistics in two extensive studies published in the early 1990s (Zarnowitz, 1992; Zarnowitz & Braun, 1992). The survey, named after its hosting institutions, namely the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER), was launched in 1968 and collected macroeconomic forecasts from on average 20 forecasters and institutions each quarter. Since 1990, the survey is conducted under the auspices of the Federal Reserve Bank of Philadelphia. Zarnowitz compares the ASA-NBER data with other prominent forecast samples, notably the data from the Research Seminar in Quantitative Economics of Michigan University, the official government forecasts from the Council of Economic Advisors (CEA) and the forecasts from the Livingston survey, the oldest compilation of (rather informal) forecasts. The MAEs of all nominal growth forecasts, calculated for each survey over the whole period, range within 0.8 and 1.2 (ASA-NBER: 1.2). The corresponding values for RGDP ($1.0 \leq \text{MAE} \leq 1.6$; ASA-NBER: 1.1) and inflation ($1.0 \leq \text{MAE} \leq 1.4$; ASA-NBER: 1.2) are slightly higher.[1]

McNees (1992) studies the forecast record of nine major forecasting institutions in the US from 1971-1991. Forecasts of nominal GNP ($1.2 \leq \text{MAE} \leq 2.1$) are found to be slightly more accurate than real growth forecasts ($1.2 \leq \text{MAE} \leq 1.6$). The inflation forecasts are clearly more accurate with an MAE between 0.5 and 0.8.

Likewise considering but the US, Joutz & Stekler (2000) examine the forecast record of the Federal Reserve Bank. Their sample consists of quarterly nominal and real GDP growth forecasts between 1965 and 1989. These forecasts are transformed into the corresponding predictions of annual growth and inflation with a one year horizon. Joutz & Stekler's findings once more confirm our general statement: The MAEs of nominal (1.68) and real growth (1.66) are clearly larger than 1% while the inflation forecast's MAE (1.16) is smaller.

Öller & Barot (2000) have studied "the accuracy of European growth and inflation forecasts". The authors' sample includes OECD forecasts of GDP growth and inflation for 13 European countries plus the corresponding forecasts of one national institute for each of these countries from 1971 to 1998.[2] Öller & Barot observe "unacceptably large" (p. 295) forecast errors: As the RMSEs vary between 1.3 and 2.6 with a mean of almost 1.9 we may infer according to our rule of thumb that the MAEs vary between $\frac{1.3}{1.25} \approx 1$ and $\frac{2.6}{1.25} \approx 2$ with a mean of about 1.5. The mean of the inflation forecast errors is slightly lower ($\text{RMSE} \approx 1.6$, thus $\text{MAE} \approx 1.3$).

A very extensive survey is conducted by Blix et al. (2001) from the Swedish Central Bank. It includes more than 52,000 forecasts from more than 250 different institutions, most of them from the private sector. The examined growth and inflation forecasts for the United States, Japan, France, Germany, Italy and Sweden cover the decade from 1991 to the year 2000. The authors find an average RMSE of 1.07 for growth and 0.61 for inflation forecasts. Again, we calculate the corresponding MAEs of $\frac{1.07}{1.25} \approx 0.9$ and $\frac{0.61}{1.25} \approx 0.5$ respectively. Notice that no study reports lower forecast errors than these ones.

Loungani (2001) extends the geographical perspective further including real GDP forecasts for 22 industrialized and 41 developing and transition-countries

1 That Zarnowitz diagnoses, in contrast to most other studies, a higher MAE of inflation than growth forecasts might be explained by the facts that (i) his sample period is relatively large and (ii) inflation forecasts are characterized by lower MAE today than 40 years ago, as we will see in the next chapter.

2 However, forecasts of several different research institutes had to be considered for some countries as no single institute covered the whole period.

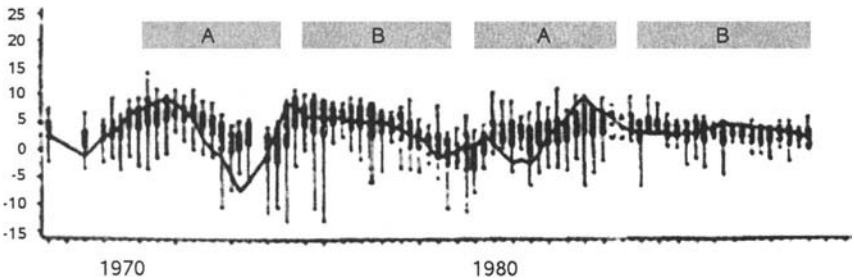



Figure 2.1: ASA-NBER record: Forecast distributions and actual percentage change of real GNP, 1968-1990. (Adapted from: Zarnowitz & Braun, 1992, p. 54)

in his sample. The sample period covers the decade from 1989 to 1998. The forecasts examined by Loungani are so-called consensus-forecasts reflecting the "mean" forecast of private sector institutions. They are provided by the same organization, namely *Consensus Forecasts Inc.*, which delivered the data for the study of Blix et al. (2001). Loungani reports mean absolute forecast errors of 1.4 for industrialized and of 2.1 for developing countries. He states that the mean growth in the industrialized countries under examination was 2.3 so that the relative MAE (the MAPE2) is $\frac{1.4}{2.3} \approx 60\%$. At a first glance, developing countries' forecasts seem to be characterized by larger errors; but when accounting for variability in the following section we will qualify this observation.

*Forecast performance in relative terms*

So far, some main studies have been introduced and their general findings have been stated in absolute terms. These results shall be qualified in the following. In fact, the MAE averaged over a long period and different institutions and countries is but a very rough rod which omits important factors that influence forecast performance. In particular, it is obvious that forecast errors depend on the specific behavior of the forecasted variable: Volatile variables are more difficult to forecast than stable ones. The volatility of major macroeconomic variables actually varies over time and we should expect that forecast performance varies with it. Figure 2.1 shows the ASA-NBER quarterly RGDP forecasts for a five quarter horizon from 1968 until 1990. While the continuous line represents the actual output, the vertical bars correspond to the forecast's statistical characteristics: Each bold bar plots the forecast sample's mean plus/minus the standard deviation, and each thin bar with its small horizontal lines indicates the whole range of forecasts in the ASA-NBER sample.

A high standard deviation of forecasts indicates a high uncertainty among the forecasters and a lower forecast precision. It inevitably results in higher forecast errors. This said, it is evident from figure 2.1 that forecast performance is much better during calm periods than in turbulent times. Consider the different periods marked by grey bars A and B. When the actual-outcome curve is steep (A) as in 1972-74, in 1980 or in 1982, the bold bars do not even touch this curve! On the other hand, in periods with rather low volatility (B), uncertainty among forecasters has decreased and the bold bars fit much better to the curve than before, i.e. forecasts are not only more precise but also more accurate. Similar conclusions hold for nominal GNP and inflation.

The lesson to be learned is that forecast evaluations should try, at least when aiming at inter-temporal comparisons, to account for this period-dependence of forecasts. The above studies use at least one of the two equivalent methods from the introduction — in short: division by standard deviation or comparison with naïve benchmark — in order to measure the relative performance of forecasts. In sum, they come to the result that macroeconomic forecasts do mostly outperform the naïve benchmark. Yet, they do so with marginal lead, and it is not always clear whether their lead is significant or not.

Here is what our authors say in some more detail: Zarnowitz (1992, pp. 399 and 526) concludes that the annual forecasts of GNP, RGNP and the implicit price deflator from the ASA-NBER survey outperform different naïve benchmarks. Whereas Joutz & Stekler (2000) and Blix et al. (2001) come to the same conclusion[3], McNees (1992) and Öller & Barot (2000) obtain more differentiated results by introducing an interesting modification in the naïve no change model for growth. They distinguish whether the basis for the no change model is (i) the output of the quarter $t$ when the forecast is made although the forecaster might not know this value and has but preliminary evidence for it or (ii) the output of the preceding quarter $t-1$. Option (i) results in a "lead"-benchmark while option (ii) yields a "lag"-benchmark. As the authors show (and as we would expect), forecasters have much more difficulty in beating the lead- than the lag-benchmark. According to McNees, less than 10% of the RGNP forecasts did not improve upon the lag-model. However, only half of them were able to beat the lead-benchmark. Decreasing the forecast horizon to half a year, more than 90% of the RGNP forecasts were outperformed by the lead-benchmark. Öller & Barot (2000) use a naïve "average change" model and a lead-model similar to McNees'. They find that a majority of 6 (out of 13) OECD and 10 (out of 13) national institute forecast series outperform the average change benchmark. However, "only a few forecasts"

3 Blix et al. find but a single forecaster, a french bank, which is outperformed by the no change benchmark.

could keep up with the lead-straw man! Loungani (2001), finally, both accounts for variability and compares the growth forecasts with a naïve no change model (on the basis of annual data). He finds that (i) when accounting for volatility the forecasts for the developing countries are slightly better than those for the industrialized countries and (ii) one-year-ahead growth forecasts are only slightly better than the naïve model; small modifications as increasing the forecast horizon by considering forecasts already released in April instead of October lead to the result that the naïve model wins!

### *The role of forecast horizon*

There is a general agreement in the forecasting literature that the absolute forecast error positively depends on the forecast horizon. Zarnowitz & Braun (1992, p. 26) show how both the RMSE and MAE of GNP, RGNP and inflation forecasts are increasing with larger forecast horizons. McNees (1992, p. 26) demonstrates that forecast performance is improving even in the course of a quarter. In addition, Joutz & Stekler, Loungani, as well as Makridakis & Hibon (2000) confirm the positive dependency. More specifically, Loungani establishes a strictly monotonic relationship not only between forecast horizon and absolute forecast error but also between forecast horizon and relative forecast error. His findings imply that growth forecasts are outperformed by a naïve no change benchmark for horizons larger than one and a half years (p. 423). However, this result may not be considered as a consensus because, as reported above, McNees (1992) found that the *relative* performance of a forecast decreases with a declining horizon. Be that as it may, since evaluated forecasts do not cover horizons larger than two years and as even one year horizon forecasts only closely outperform naïve benchmarks, we may infer that economic practitioners and academics agree that macroeconomic forecasts cannot reasonably be derived for horizons larger than two years. And this is a conservative estimation.

### *Directional forecasts: predicting turning points*

Hitherto, we have studied the performance of point forecasts. It might, however, be the case that trend forecasts perform much better than their non-directional counterparts. But how can we assess directional forecast performance? One way to do so is by observing whether forecasts succeed in predicting turning points, i.e. changes in direction. Now, the empirical literature agrees overwhelmingly that trend forecasts consistently fail according to this criterion. Almost every study on macroeconomic forecast evaluation mentions the failure of these forecasts to predict turning points[4] and no study at all reports contrary evidence. Let us have a closer look at some of the findings.

Öller & Barot ask whether macroeconomic forecasts indicate at least the correct sign of next year's growth, i.e. whether they "issue correctly timed signals". And they continue, "The sad answer is, only in rare cases". (p. 306) This conclusion is inferred by comparing the GDP forecast once again with a naïve no change benchmark of the (easier-to-beat) lag-type.[5] How well do the professional forecasts of the OECD and the national institutes perform against this straw man? In eight out of 26 cases, the professional forecasts were outperformed by the naïve benchmark! In addition, half of the 18 remaining professional forecasts are not correlated with actual growth. Öller & Barot conclude that "this contradicts the claim [...] that macroeconomic forecasters may be better at direction than at numerical accuracy" (p. 307).

Loungani examines the specific forecast performance during recessions. He identifies a total of 60 recessions (defined as a year in which real GDP declined) in his sample period and finds that the consensus forecast has predicted negative growth in only three of these 60 cases (p. 424). Thus, in 57 cases ($\approx$95%) positive growth had been predicted when negative growth actually occurred! This has to be considered as another strong evidence for directional forecasts' failure.

## 2.2 The performance of different forecasters ...

Having studied the overall performance of macroeconomic forecasts so far, we will now turn our attention to the forecasters: Are there any differences in their forecast performances? A rough categorization distinguishes three types of forecasting institutions: (a) official, (b) academic, and (c) private forecasters. Official forecasters include international organizations as for instance the International Monetary Fund (IMF) or the World Bank, central banks, government agencies or advisors like the CEA. The category of academic forecasters is made up of national research institutes and universities. Finally, there is the large industry of professional forecasting services. Two prominent surveys of private forecasters in banks, investment fonds and other firms are those by *Blue Chip* and *Consensus Forecast Inc.*

These are the general findings of the empirical studies: First of all, the vast majority of studies confirms that there exists either only a small or no difference at all among the forecasters with regard to their forecasting performance; no study mentions negative evidence on this point. In line with this, no single forecaster

4 Namely: Blix et al. (2001), Burns (2001), Loungani (2001), Joutz & Stekler (2000), McNees (1992), Osborn et al. (2001), Öller & Barot (2000) and Zarnowitz (1992).

5 Notice that such a naïve no change benchmark cannot predict turning points at all!

can be considered as the best performing one. Finally, it is controversial whether pooling forecasts improves performance. But notably recent studies suggest that this is not the case.

As to the first point, McNees (1992), examining all three types of forecasters, finds that there are small differences among the best performing institutions. In addition, no single forecaster outperforms the other institutions consistently in all variables and over the whole sample period. In contrast but not in contradiction to this dense winner-field, there are some forecasters who constantly perform badly. In agreement with McNees, Öller & Barot (2000) find that official (OECD) forecasts are very close to those released by national research institutes. An almost perfect correspondence between private sector consensus forecasts and institutional forecasts by the IMF, the OECD and the World Bank is observed by Loungani (2001). Furthermore, the Fed's forecasts do not differ significantly from the ASA-NBER survey according to Joutz & Stekler (2000). Blix et al. (2001) find strong similarities between forecast errors of academic, official and private forecasters, too. They interpret this finding as indicating a herd-behavior of forecasters who tend to make the same forecasts as well as the same revisions. This conclusion is also shared by Coyle (2001), former organizer of *The Independent*'s "Golden Guru" forecasting competition.

Given the rather similar performance of forecasters, does pooling yield better forecasts? According to Zarnowitz & Braun (1992), it is widely accepted that joining individual forecasts together into one consensus forecast (calculating the average forecast) increases the forecast performance. Their survey of the ASA-NBER data for the 1970s and 80s reconfirms this statement. More recently, this view has been shared by Öller & Barot: While the authors find that some of the studied forecasts do not outperform a naïve benchmark, they concede that pooling the forecasts of the OECD and the national institutes results in a group forecast that clearly outperforms the straw man. On the other hand, Blix et al. (2001) find no improvement comparing the group forecast error with the mean error of the individual forecasts.

How can these contradicting observations be consolidated? Consider some forecasts with an error sequence $\langle e_1, \ldots, e_n \rangle$. The error of the corresponding consensus forecast is accordingly,

$$\overline{e} = \frac{1}{n} \sum e_i.$$

Now, it follows immediately that the MAE (see section 1.2) of the error sequence is greater than or equal to the MAE of the consensus forecast's error which is in fact simply $|\overline{e}|$. As the RMSE is generally greater than the MAE, consensus forecasts are characterized by lower overall errors than the original forecast sequences. Yet, if herding occurs, the individual forecast errors have the same sign. In that case, the MAEs of the forecast sequence and the consensus forecast are equal. According to our rule of thumb, the RMSE of the forecast sequence thus is about a quarter larger than the consensus forecast's RMSE. Hence, pooling does not improve forecast performance as measured by the MAE (and only slightly improves it as measured by the RMSE) under herd-behavior. If herding had thus become a general practice during the 1990s, the contradicting empirical results would be reconciled. And indeed, Kacapyr (1996) reports that it has become a common strategy not to rely on one single source alone in producing a forecast. This development might even have been encouraged, if not triggered by the finding that pooling increases performance. But as soon as sufficiently strong herding had occurred, pooling turned out to be of no use anymore.

## 2.3 ... and different forecasting methods

In this section, we will have a closer look at different forecasting methods and their specific performance. Practitioners indeed use a large variety of different methods to generate macroeconomic forecasts which range from personal judgment and simple extrapolation to forecasting with sophisticated models incorporating hundreds of equations. As far as I can tell, there is no canonical taxonomy of forecasting methods; quite the opposite, we find fairly different categorizations in the literature. Kacapyr (1996) for instance distinguishes three main methods: indicator based forecasting, time series methods and econometric forecasting. Zarnowitz (1992) adds to these the survey based forecasting. Clements & Hendry (2002*b*), however, list seven different methods: informal forecasting, expert judgment, extrapolation methods, indicator based forecasting, survey methods, time series methods and, finally, econometric modeling. These categories, however, are not mutually exclusive: Extrapolation can be considered as a subcategory of time series models; surveys that yield business climate indices have a lot in common with other indicators as for example stock prices. Expert judgment, as long as not only a single expert is concerned, prerequisites a survey, etc. In the following, I attempt to give a more structured overview of economic forecasting methods.

The dichotomy of predictive data and predictive inference (see p. 9 above) provides a matrix to categorize forecasting methods. Economic forecasts of a variable $X$ can be based on two types of data: the past values of $X$ on the one hand, and these values plus the past values of other (economic) variables, including survey data, on the other hand. With regard to the different inferences used for economic forecasting, we can distinguish two main categories: in the first place, inferences

| | *Method of inference* | | |
|---|---|---|---|
| | Pattern extrapolation | | Law based |
| *Data* | Informal | Formal | |
| Past record of predicted variable | Charting | Univariate time series models | – |
| Arbitrary economic variables | Indicators and economic surveys | Multivariate time series models | Econometric models |

Table 2.2: Matrix of economic forecasting methods.

based on economic theory, and, in the second place, methods that do not rely on economic theory like pattern recognition and extrapolation. This second category is subdivided into formal inferences based on statistical methods and informal, heuristic inferences. Table 2.2 plots the resulting matrix and fills it with the methods mentioned earlier. These methods shall be introduced briefly in the following paragraphs before their specific forecasting performance will be compared.

*Time series models*

*Univariate* time series models use the record of past values of a certain variable to predict its future values: Let $X$ be the variable to be predicted, $(X_1, \ldots, X_n)$ its sample record, and $k$ some number. The crucial assumption states that the value of $X$ at time $t$ is a function of its last $k$ values,

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-k}, a_1, \ldots, a_m), \tag{2.1}$$

where $(a_1, \ldots, a_m)$ is a set of parameters and $k$ is the so-called lag-term. By application of statistical techniques, the parameters are calculated in such a way that the equation (2.1) fits the data sample $(X_1, \ldots, X_n)$. The fully specified equation can then be used to predict the future values ${}_nX^F_{n+1}$ and, recursively, ${}_nX^F_{n+2}, {}_nX^F_{n+3}, \ldots$.

*Multivariate* time series models do not only use the predicted variable's own record, but also that of other economic variables to derive forecasts. Formally, the value $X_t$ is supposed to depend on its past record plus the past values of $l$ other variables $X^*, X^{**}, \ldots, X^{l*}$. Defining $\vec{X}_t := (X^*_t, X^{**}_t, \ldots, X^{l*}_t)$, the multivariate equivalent to (2.1) is,

$$\vec{X}_t = f(\vec{X}_{t-1}, \vec{X}_{t-2}, \ldots, \vec{X}_{t-k}, a_1, \ldots, a_m). \tag{2.2}$$

As for the univariate case, one can specify the parameters $a_i$ via statistical methods. The future values of $\vec{X}$ can then be obtained recursively in analogy to the univariate case. In contrast to the latter, this multivariate method not only yields the forecast of the predicted variable $X$, but also of all the other variables included in the model. They are co-predicted.

The plurality of time series models arises from the forecaster's freedom to chose the lag-term, the set of other variables to be included and finally the specific form of the functional relation, i.e. whether $f$ is a linear function, a polynomial, an exponential function, logarithmic, periodic or whatever. The forecaster typically improves his time-series model via trial and error.

This said, time series methods obviously do not rely on economic theory.

*Charts, Indicators and Surveys*

Although charting on the one hand and economic forecasting with indicators on the other hand are two different methods, they resemble each other in many aspects and face similar problems. For this reason I present them jointly. Charting refers to a method used by financial analysts with the objective of forecasting the future development of financial variables, notably stock market prices. Chartists, as these analysts are called, plot past financial data in a diagram and then try to identify and recognize well known patterns in the chart that can be extrapolated into the future. Malkiel (1999) gives an amusing introduction into the art of chartism. What follows is based on his account.

Have a look at figure 2.2: This is the typical way in which chartists draw prices of a single stock. Each unit on the x-axis represents a trading day while the y-axis measures the stock price. A vertical line indicates maximum and minimum prices of the stock while the small horizontal line represents the final price at that specific day. Now, we see a quite regular pattern in this diagram, it is the most famous pattern of chartism at all: the so-called head-and-shoulder formation. This pattern resembles a human torso with the two exterior bows interpreted as shoulders and the major bow in the middle as the head. An important feature of this pattern is that the curve slightly touches the "neckline" several times but never falls below it. The chart forecast states: (i) If the curve crosses the neckline after having finished the head-and-shoulder formation, then the stock price will continue to fall. Or, with the metaphors of the stock market, crossing the neckline is a bearish signal. But the bear can be trapped! If the forecast (i) fails and the price increases dramatically although having crossed the neckline, chartists predict (ii) an enduring rise of the particular stock price as depicted in figure 2.3.

This example illustrates that chartists base their prediction on the same kind of data which is used in univariate time series models, i.e. the past record of the

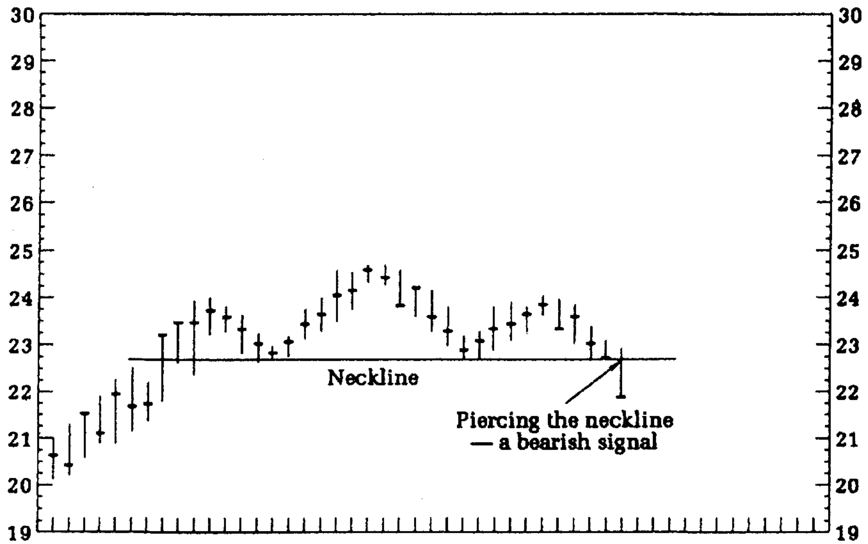



Figure 2.2: Head-and-shoulder formation: predicting falling prices. (Source: Malkiel, 1999, p. 122)

predicted variable. But chartists try to discover typical patterns in the graphical representation of the data with their own eyes! There are some intrinsic flaws of this method. They all relate to the problem of data interpretation: As real charts rarely are as clear as ours, what still counts as a head-and-shoulder formation, and what does not? Data might be ambiguous so that different patterns with contradicting implications could be recognized in the chart. Charting seems to have the favor of clairvoyance with chartists 'seeing' future in typical patterns occurring in chaotic processes.

Nowadays, chartists do no longer have to draw charts: They rely on computer power. In fact, the idea of pattern recognition in a time series is perfectly realized in a time series model. Hence, chartists became what is known as technical analysts, checking share price series with computer programs for underlying patterns. Although time series analysis does not have the above mentioned problems of chartism, only a small minority of financial analysts in fact use technical analysis. By the end of chapter 7, we will see why.

The macroeconomic counterpart of charting is the wave theory of business cycles (see Kacapyr, 1996). Wave theorists interpret business cycles as the interference of enduring waves with different wave length. Similar to astrologers, they explain severe recessions or depressions as "conjunctions" of the waves' valleys. But macroeconomic forecasts do not have much to do with this. However,

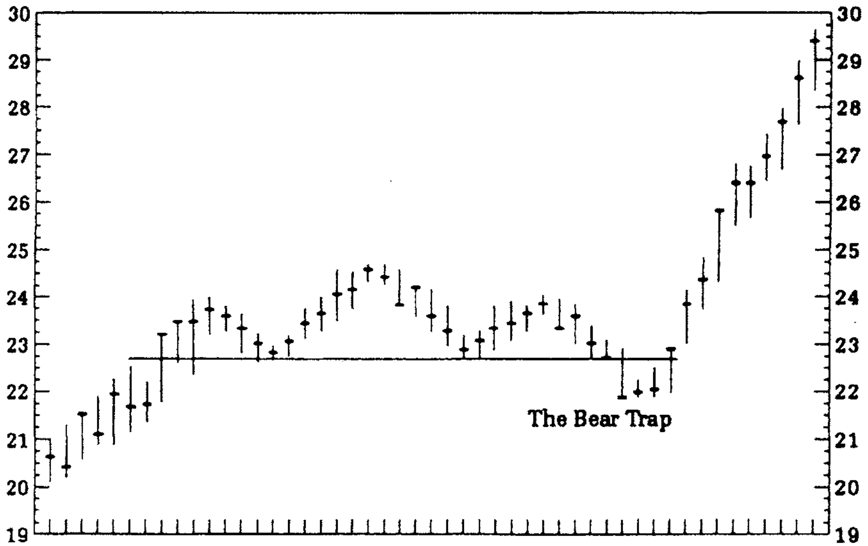



Figure 2.3: Bear trap: promising further share price rise. (Source: Malkiel, 1999, p. 123)

charts still have a role to play. In opposition to financial charting, economists do not plot GDP charts but charts of other economic variables that are thought to "lead" growth. Forecasting with indicators involves the idea that at least one leading variable anticipates the movement of the variable to be predicted. This does not necessarily require a causal relation. The monthly number of demands for building-permits might for instance represent a leading variable for growth. But demanding a permit clearly does not cause economic growth.

A famous economic indicator system is the leading economic indicators index from the Conference Board, compiled by the US Department of Commerce (see The Conference Board, 2003). It is supplemented by an index of coinciding and lagging variables whose movements coincide with or, respectively, follow the growth dynamic. Since the indicator method is basically a graphical, non-quantitative analysis, it faces the same problems as chartism. Kacapyr (1996) cites as its general problems the interpretation of ambiguous movements, in particular the question when a change in the leading indicator is *substantive* so that a forecast may be derived. Although different rules for interpreting indicators have been elaborated, none has resolved these problems in a satisfying way. A final problem emerges according to Kacapyr because the multitude of indicators actually applied will hardly lead to consistent but rather to contradicting forecasts.

Some economic forecasters conduct surveys in order to assess the consumer or business climate.[6] Since the expectations of economic agents influence their behavior, i.e. their investment or consumption decisions, economists hope to extract useful information from such survey data. Yet these business or consumer climate surveys can be considered as another version of the indicator method, and they do indeed face the same problems.

*Econometric models*

In contrast to the forecasting methods introduced so far, econometric methods rely on economic theory. Here comes an example of how to construct an econometric model and how to use it for forecasting: The following equations define a small mathematical macroeconomic model,

$$\begin{aligned} I &= I(Y,K) \\ S &= S(Y) \\ \dot{Y} &= \alpha(I-S) \\ \dot{K} &= I(Y,K) - \delta K \end{aligned} \tag{2.3}$$

Translated to everyday English, these equations state that (i) the investment $I$ depends on the total production $Y$, i.e. the GDP, and the capital stock $K$, (ii) the total private savings $S$ are a function of the GDP, (iii) the economic growth ($\dot{X}$ being a short cut for $\frac{dX}{dt}$) is determined by the difference between investment and savings multiplied by a parameter $\alpha$, and finally (iv) the change in capital stock equals the difference between investment and capital deterioration where $\delta$ is the deterioration rate. These functional dependencies are further specified in economic theory by statements relating to the different derivations. In fact, (2.3) is hardly an economic model without such specifications but we will skip them here for the sake of presentation and because we will encounter a particular instance of this model-type in chapter 8.

The above equation system is highly abstract. The theoretical model can be transformed into an econometric model by specifying the functional relationships including the introduction of parameters so that the left-hand side of each relationship could in principle be calculated from its right-hand side. Furthermore, an error term — a random variable — is attributed to each relationship. We conse-

6 As for example the monthly survey conducted by the European Commission.

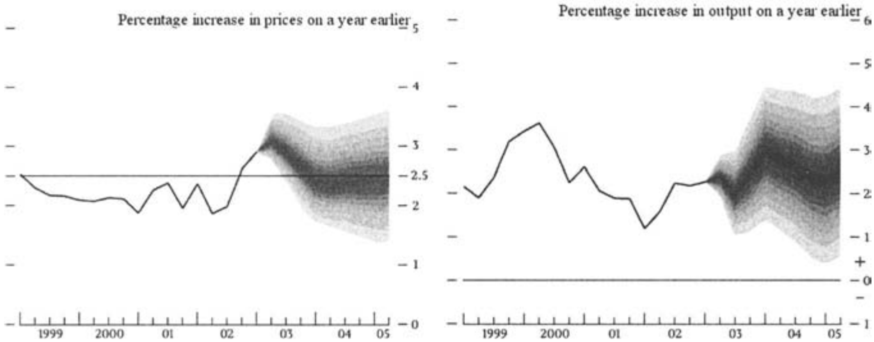



Figure 2.4: Predicting probability distributions: The Bank of England's fan charts for inflation and growth. (Source: Bank of England, May 2003, p. iii)

quently obtain,

$$\begin{aligned} I &= I(Y, K, a_1, \ldots, a_m) + u_I \\ S &= S(Y, b_1, \ldots, b_n) + u_S \\ \dot{Y} &= \alpha(I - S) + u_Y \\ \dot{K} &= I(Y, K) - \delta K + u_K \end{aligned} \tag{2.4}$$

This econometric model with parameters $(a_1, \ldots, a_m, b_1, \ldots, b_n, \alpha, \delta)$ and the error terms is then fitted to the data, a sample of past values of $I$, $S$, $Y$ and $K$. In applying the same statistical methods as used in time series models (see above) numerical values of the parameters and the stochastic properties of the error terms, namely their mean and variance, can be obtained. The numerically specified model finally renders it possible to calculate future values of the endogenous variables, i.e. to make forecasts.

In fact, econometric methods not only allow for deriving point forecasts but also density forecasts because of the introduced random variables. This procedure is for example implemented by the Bank of England as illustrated by figure 2.4 showing its inflation and growth forecasts. However, such probability density forecasts will not be considered in this part because there are no results of empirical evaluations to report. Besides the fact that only a few forecasters have recently started to issue density forecasts, their evaluation is very tricky for systematic reasons.[7] The credibility of economic probability forecasts will therefore be appraised indirectly in chapter 11 given the performance of non-probabilistic forecasts and its explanations.

It should be clear from what has been said so far that econometric methods, when used for forecasting, require extensive calculations. These are commonly performed by computers and that is the reason why the rise of econometric fore-

[7] In a single sentence: The problem is that probability forecasts cannot be strictly falsified.

casting coincided with the rise of information technology. Computer power allows for calculations largely beyond those implied by our simple model. In fact, well-known econometric models such as the DRI/McGraw-Hill model incorporate more than a thousand equations specifying functional relationships between economic variables.

Before comparing the forecasting performance of the different methods, we should briefly discuss another characteristic feature of econometric models. Econometricians typically stress that econometric models can not only be used for categorical but also for conditional forecasting (for example Turner, 2001; Barrell, 2001). This is easily understood if we consider the distinction between endogenous and exogenous variables.[8] All the variables of our example model are endogenous. Now, assume that the change of GDP not only depends on the difference between saving and investment but also on government spending, $G$. An equation accordingly modified for economic growth for instance is,

$$\dot{Y} = \alpha(I - S) + g \cdot G,$$

where $g$ is a further parameter. Since the modified model includes no equation determining the future values of $G$, we cannot calculate any future values (of the endogenous variables) without co-predicting public spending! This cuts both ways: Forecasting $G$ makes the categorical forecast more reliant on further explicit assumptions. On the other hand, we can now generate conditional forecasts predicting what would happen *if* government spending were increased over the next 10 years, or *if* government spending did not change, or ...

*Forecasting performance of methods*

There are two different ways to assess and compare the performance of forecasting methods to which I will refer as the laboratory- and the field-test. The laboratory-test consists in collecting (i) a large sample of time-series and (ii) a large sample of forecasting methods in order to apply the methods to the different time-series. The errors of each method are averaged and yield the method ranking. This procedure is also known as forecasting competition. In contrast to the laboratory-approach, field-tests require to collect published forecasts and to categorize them according to their underlying method. While this second procedure usually suffers from incomplete samples it has the advantage of being based on real forecasts. In the following, I will present results from tests of both types.

[8] For a given model, a variable is said to be endogenous if and only if its values are predetermined by the model once an initial state is given. If a variable's value is not determined by the model but nevertheless included in one or more equations, this variable is called exogenous.

Forecasting competitions have been carried out since the late 1960s. Reid (1969) and Newbold & Granger (1974) performed the earliest competitions involving different time series methods. The best known competitions, however, probably are the so-called M-Competitions named after their initiator Spyros Makridakis. The first M-Competition dates back to 1982; Makridakis compared 15 different predictive methods over a total of 1001 micro- and macroeconomic time series. The calculations were performed by forecast specialists. The second M-Competition was launched at the beginning of the 1990s and was, in contrast to M1, a real-time competition. This means that participants did not generate out-of-sample forecasts of past, already observed variables but they really had to predict future values. The six macroeconomic time series were taken from the US. Finally, the third M-Competition took place in 1999 and included 3003 different time series, among them more than 700 macroeconomic time series. A total of 24 predictive methods were evaluated. The M-Competitions compared *time series* methods only, but the sample of M3 included such different methods as extrapolation and univariate time series methods on the one hand and neural networks as well as sophisticated expert systems that automatically pick a time series method given the sample data on the other hand.

The M-Competitions gave rise to plain results which have been reproduced by the three competitions and confirmed by several other studies as Fildes & Ord (2002) and Makridakis & Hibon (2000) report. The following three conclusions are of special interest to us: (1) Statistically sophisticated models perform as well as simple ones. Complexity (of the method) does not necessarily lead to more accuracy, and the difference between the methods' errors was generally small. (2) Combining different forecast methods yields a lower error than the average individual method. (3) The ranking of the methods depends upon the chosen forecast measure.

These results suggest to draw an analogy between the performance of forecasters and methods: There are small differences both between methods and forecasters. And pooling individual forecasts (necessarily) improves the performance as long as there is no herding-behavior — which is quite rare among mechanically applied methods. Still, before we jump to conclusions, we should consider other forecasting methods as well.

Carrying out a field-test, Armstrong (1984) confirms the results from the preceding paragraphs. He reports findings on the forecast performance of econometric methods compared to formal or informal extrapolation methods from a dozen of studies mainly dating from the 1970s. Armstrong concludes (i) that no study established a *significant* superiority of econometric models over time series methods and (ii) that econometric models do not even tend to be superior since there were as many studies that found econometric models to be more accurate as there were studies that came to the opposite result. Another question raised in that paper is whether complex econometric models outperform less complex ones. Including also evidence from other social sciences than economics in his sample, Armstrong essentially confirms the M-Competitions' conclusion: There is no *significant* superiority of complex models over simple ones. Here, too, Armstrong faced conflicting, out-balancing evidence.

The ASA-NBER survey not only contains forecast series but also information about the forecasting methods that were used to produce the forecasts. The survey distinguishes four methods: leading indicators, anticipations surveys, econometric models with the two subcategories own and third-party models, and informal models. Participants of the survey had to indicate which methods they used and were asked to rank them according to their importance for the forecast generation. Zarnowitz (1992, pp. 403ff) analyzes this information for three short time intervals: 1968-70, 1974-75 and 1980-81. Grouping the forecasts according to the most important method they are based on, he finds that the RMSEs characterizing the different methods range from 0.89 to 1.09 with third-party econometric models yielding the best and 'own' econometric models the worst error. The other methods lie almost exactly in between those two. In addition, Zarnowitz finds that the picture does not change essentially for RGNP or inflation forecasts. For the period from 1970 to 1985, Zarnowitz (1992, p. 528) compares the forecasts based on major econometric models (as for instance those by the Wharton Econometric Forecasting Association or DRI/McGraw-Hill) on the one side with sophisticated formal time series models on the other in a mixed laboratory- and field-test. He concludes that both perform equally well.

Camba-Mendez et al. (2002) evaluate the indicator-system used by the OECD. Different sophisticated indicator models are compared with a univariate time series model and a no change benchmark. The authors find that the different indicator forecasts do not perform any better than the simple univariate time series model. Furthermore, the indicator forecasts are, in the short term, largely outperformed by the naïve straw man.

## 2.4 Forecasting financial variables

Forecasting financial variables could be an extremely profitable business. Yet, forecasting exchange rates or stock markets is much more difficult than predicting growth or inflation. Financial forecasts are commonly evaluated against a random walk benchmark. Such a random walk model (without drift) is in fact nothing but our well-known naïve no change model. For if a variable follows a random walk, the best prediction we can make is the no change forecast. In the early 1980s, Meese & Rogo (1983*a*,*b*) evaluated different exchange rate models that were based on "fundamentals" (such as GNP or interest rate) against the random walk model and concluded that no model consistently and significantly outperformed the random walk. This result was a shock for the financial economics community as Clarida et al. (2001) and Meese & Rogoff (1988) report. It has become known as the random walk hypothesis and it was widely accepted in the following two decades.

In a recent study that reexamines the random walk hypothesis Degrér et al. (2001) evaluate 30 different exchange rate models which predict the exchange rate of the Swedish Krona for forecast horizons up to 12 quarters and which are based on fundamentals as GDP, inflation, interest rate, budget deficits etc. Now, these fundamentals are exogenous variables so that, as we have seen above, co-predictions of these fundamentals are required in order to derive the exchange rate forecast. The exchange rate models can hence be evaluated retrospectively under two assumptions: (i) perfect foresight, i.e. the exchange rate forecast is calculated using the a posteriori known actual outcome of the fundamentals as co-prediction or (ii) best available forecast, i.e. the exchange rate forecast is calculated using the predictions of the fundamentals available at that time. The authors find that even under idealistic assumption (i) only a few models beat the naïve straw man. Under the second assumption, *none* of the models under examination do so! Degrér et al. conclude, reaffirming Meese & Rogo (1983*a*): "None of the models could convincingly outperform a random walk alternative." (p. 10)

Mariano (2002) reports the results of an interesting (laboratory) evaluation of stock market predictions by Sullivan et al. (1999). The authors altogether examine 8,000 different trading rules taken from the literature on stock market trading or from previous similar studies. Such trading rules implicitly rely on forecasts of stock prices and yield the more profit, the more successful the underlying forecast. The sample period the trading rules have been calibrated to covers 90 years from 1897 to 1986. For this sample period, the best trading rule delivered significantly higher returns than a naïve straw man: a simple buy and hold strategy. In other words, in-sample forecasts outperformed the naïve benchmark. Yet, the litmus test is their out-of-sample performance in the following decade (1986-97). The authors find "no evidence of superior performance of the best trading rule" (cited from Mariano, 2002). Malkiel (1999), in an engaging plea, joins this position. However, he does not bother examining thousands of different trading rules but, in what I called a field-test, simply compares the performance of professionally managed equity funds that apply a mixture of these trading rules with the overall development of the stock market: Even without accounting for trading fees, the average equity fund is outperformed by the simple buy and hold strategy.

The random walk hypothesis has been the subject of different critiques. Mark (1995) was one of the first to undermine the random walk hypothesis in showing that although exchange rates are not predictable in the short run, models based on fundamentals may outperform a random walk at longer horizons. Meanwhile, economists have tried to improve their forecasts by the introduction of nonlinearities into the models. But this did not necessarily improve their performance: Cao & Soofi (1999) found no significant superiority of non-linear models compared to a random walk. However, a recent study claims to have beaten the random walk model consistently and significantly for the first time: Clarida et al. (2001) exploit the internal structure of the forward exchange rate, i.e. the exchange rate set today for a foreign currency transaction with payment or delivery at some future date, and apply a non-linear model to produce short-term out-of-sample retrodictions of spot exchange rates of different currencies. Although the superiority of the model compared to the random walk is striking, one retrospective study hardly establishes the claim that financial forecasts succeed *consistently*. It still seems to me an open question — to be further discussed in chapter 7 — whether that evidence even indicates financial forecast credibility.

## 2.5 Common problems

Empirical forecast evaluations face typical problems of which we have already encountered some. These are clearly perceived and described by economists and forecasters themselves, as for instance by Öller & Barot (2000) or by McNees (1992). Before stating the robust observations that can be drawn from the discussed studies, I shall give a short synopsis of such common problems.

First of all, there is the variety of error measures. Not only do different error measures complicate the comparison of empirical studies, but the results obtained from the forecast evaluation may vary with the specific measure that is chosen.

But even if a measure is agreed upon, economists risk to get lost in dimensionality when evaluating forecasts because forecast performance seems to depend on many different factors: such as the predicted economic variable, the forecaster, the forecast horizon, the time period when the forecast is made, and the choice of what counts as actual outcome (forecasts tend for example to be more accurate for preliminary than for revised data).

It is impossible to account for all these factors in the course of the evaluation because the available data, i.e. the forecast sample, is too limited. If, for example, we aim at identifying the role of the forecast horizon, we should ideally compare forecasts that differ in forecast horizon *only* and are identical with respect to forecaster, time period, actual outcome, method applied, etc. But this sub-sample is de facto too small, if not empty, to give rise to statistically meaningful results. Any forecast evaluation thus runs the risk of comparing apples and oranges and should be judged with care. This said, we can now restate the main observations from this chapter.

## Main observations

**Observation 1 (Absolute forecast error)** *The MAE (RMSE) of macroeconomic growth forecasts is typically slightly larger (substantially larger) than 1%. Inflation forecasts are characterized by smaller errors.*

**Observation 2 (Naïve benchmark)** *Macroeconomic forecasts closely outperform naïve benchmarks. This is in general not the case for financial forecasts.*

**Observation 3 (Directional forecasts)** *Directional macroeconomic forecasts do not perform any better than point forecasts. This holds in particular because of the complete failure to predict turning-points.*

**Observation 4 (Forecast horizon)** *Forecast performance decreases in absolute and relative terms with the forecast horizon. The economic future beyond a two-year horizon (at most) has to be considered as unpredictable.*

**Observation 5 (Forecasters)** *No single professional forecasting institution consistently outperforms all the rest of the forecasting institutions (including private sector, government, academics).*

**Observation 6 (Methods)** *There are no significant differences between the different forecasting methods with regard to their predictive performance. Complex methods perform as well as less sophisticated ones. 'Theory-free' methods are as accurate and precise as methods based on economic theory.*

**Observation 7 (Problem of dimensionality)** *Forecast evaluations must be judged with care as they are tainted with problems that stem from the tension between the high number of factors forecast performance actually depends on and the comparatively small forecast record.*

# Chapter 3

# A historical perspective on economic forecast performance


## Summary

This chapter assesses the performance of macroeconomic forecasts from a historical point of view. After briefly outlining the history of economic forecasting, it reviews and discusses empirical studies that consider whether economic forecast performance has been improved over time. In the light of these empirical studies, the answer to that question is negative. Likewise, there seems to be no predictive progress as to the different forecasting methods, either. The last section summarizes this chapter's main observations.


## 3.1 The history of economic forecasting

A general history of economic forecasting remains to be written. This section does not pretend to be more than a rough and certainly incomplete sketch of what could be told as a rich and exciting story. But luckily, there is at least Mary Morgan's book reconstructing the history of econometric ideas (Morgan, 1990) which in doing so necessarily touches the history of forecasting from time to time: What follows owes a great deal to it.[1]

As Zarnowitz & Braun (1992, p. 8) tell us, the oldest records of economic forecasts date back to the 1830s and consist of directional forecasts of business activities; they have more recently been used to reconstruct business cycles. Since statistical methods like regression and correlation analysis were developed only in the last decades of the 19th century, those early forecasts must have been based on judgment, informal reasoning or simple extrapolation.

[1] Whenever mentioning a historical fact below without giving an explicit reference, I took it from Morgan's account.

The rise of statistics brought about the first attempts to deal with business cycles in a strictly quantitative way. William Stanley Jevons and Clément Juglar should be considered as the first pioneers of econometrics whose work eventually gave rise to quantitative economic forecasting. While Juglar used statistical methods in order to identify economic cycles (based on financial data), Jevons developed his own theory of business cycles according to which they are caused by sunspot cycles. As we learn from Morgan (1990, p. 23), Jevons used his theory to predict turning points on several occasions.

Henry Ludwell Moore followed Jevons when developing his own theory of the business cycle in the early 20th century: While his first version puts down business cycles to weather cycles, a later version linked these cycles to the movement of Venus. In spite of the empirical inadequacy of their theories, Jevons and Moore have to be credited for the further development and implementation of the young statistical methods.

In contrast to Jevons and Moore, and more in line with Juglar, Wesley C. Mitchell and Warren M. Persons focused on exact (statistical) description of the business cycles and were reluctant to developing *theories* of the business cycle. They attempted to set up measures of (macro)economic activity. Interestingly, the first business indicators serving this purpose had been commercial ones and emerged at the turn of the centuries in the US[2]; the *Brookmire Economic Service* and the *Babson Statistical Organization* had been pioneers of the private forecasting business (Wagemann, 1928, p. 6). Simple business barometers had already been used for economic forecasting purposes and were commercially successful by 1913.[3] On a scientific level, however, it was the research of Persons in the subsequent years that stipulated further work.

In 1917 and under the direction of Persons and Charles J. Bullock, Harvard University appointed the Committee on Economic Research and charged it with the task to develop a measure for the overall state of the economy. With main contributions by Persons who used statistical time series methods for the first time, the Harvard Business Barometer was constructed. It consisted of three indicators — a leading, a lagging and an actual variable. Those variables are (A) speculation, (B) pig-iron production plus a set of commodity prices and (C) financial markets where (A) is the lead- and (C) the lag-variable. Figure 3.1 is a historical plot of the barometer.

Since the Harvard ABC contained a leading variable, it was suited for economic forecasting. In fact, economic forecasting based on business barometers boomed in the Golden Twenties. Irving Fisher reported in 1925 that more than

[2] This agrees with a remark by Morgenstern (1928, p. 1) who dates the emergence of macroeconomic forecasting back to the beginning of the 20th century.

[3] Morgan (1990, p. 47) mentions this fact when discussing Mitchell's *Business Cycles*.

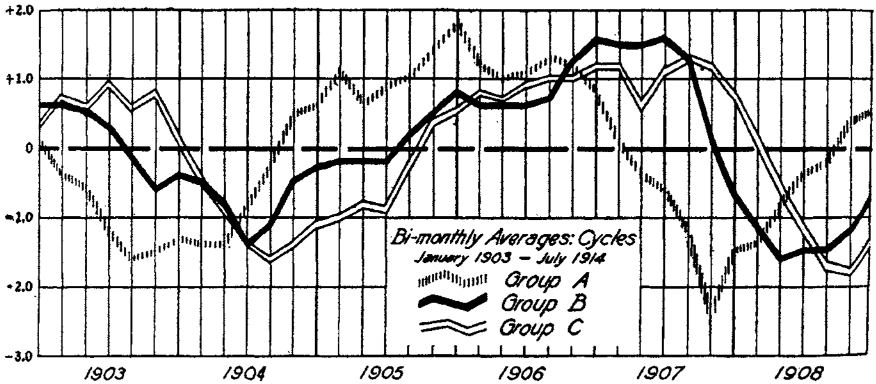



Figure 3.1: The Harvard ABC: a tripartite economic indicator system. (Source: Morgan, 1990, p. 62)

40 commercial agencies which specialized on business forecasting existed in the US (Fisher, 1925). An early evaluation of different monthly forecasts from 1929 certifies even a "modest success" (Zarnowitz & Braun, 1992, p. 10) and institutes devoted to business cycle research and the development of indicator systems spread all over Europe. In Germany, for instance, Ernst Wagemann, director of the Institut für Konjunkturforschung founded in 1925, developed a sophisticated indicator system that was supposed to yield three-month forecasts (Wagemann, 1928). Figure 3.2 is the German counterpart to the Harvard ABC; the similarity jumps to one's eyes. It is worth noticing that the forecasts generated with these business barometers were directional forecasts (Wagemann, 1927) what actually fits in our discussion of indicator based forecasting in the preceding chapter.

However, not all economists shared the optimism of their colleagues with regard to the economy's predictability. The business cycle pioneer Mitchell or the Dutch Central Bureau of Statistics, for instance, remained skeptic, or even refused any forecasting attempts at all (see van den Bogaard, 1999). This skepticism was finally confirmed by the complete failure to forecast the great depression — a breakdown that echoed in the reluctance of major economists to use their models for forecasting purposes.

The failure to predict the great depression as well as the rise of new methods caused the decline of indicator systems which had had their golden times in the 1920s. However, the modern economic indicators discussed in the previous chapter are direct descendants of the inter-war period's business barometers.

It was the emerging science of econometrics which provided new predictive methods in the 1930s. In developing the first econometric models, Ragnar Frisch and, inspired by the former, Jan Tinbergen did not restrict themselves to merely

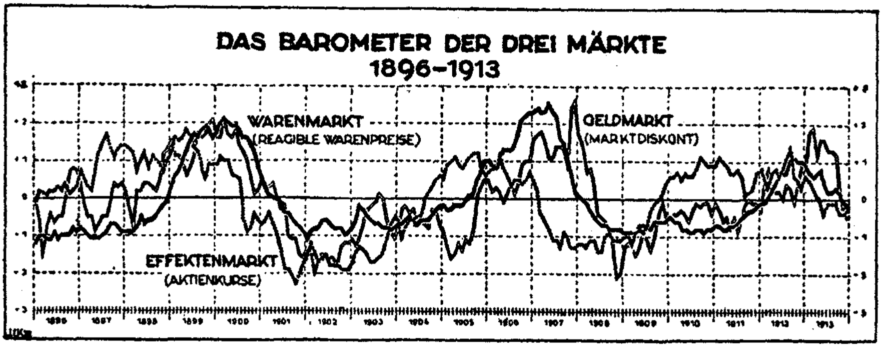



Figure 3.2: German Business Barometer: twinning of the Harvard ABC. The figure displays indicators for the years 1896-1913. (Source: Wagemann, 1927)

describing the economy via statistical methods, but tried to find formal relationships that characterize the economy as a whole and that could explain its empirical properties, in short: They set up mathematical models that were meant to be empirically adequate.

Tinbergen, working at the Dutch Central Bureau of Statistics and probably influenced by the bad experience with indicator-based forecasting, focused on policy evaluation and hence conditional instead of categorical short-term forecasting. Yet, in his 2nd report to the League of Nations from 1939, Tinbergen evaluated the predictive performance of different econometric models which varied in their investment function and covered the period between 1919 and 1932. For this purpose, he built an econometric model composed of 71 variables and 48 equations. Milton Friedman evaluated the final model of Tinbergen's report and found its predictive performance for a five-year horizon to be "unimpressive"[4].

In addition to the rise of econometrics, another development of importance for economic forecasting took place in the inter-war period: the establishment of national income accounts. Estimates of total national income date back to the late 17th century, but despite a constant improvement, such estimates had been fragmentary and insufficiently detailed for theoretical analysis for more than 200 years (Kendrick, 1970). Although the early business barometers were meant to compensate this lack of measures for the whole economy, they had to rely on selected variables that did not directly gauge the national economic output. This had been one important source of skepticism regarding the business barometers of the 1920s (van den Bogaard, 1999). Now, in the inter-war period, countries started to establish continuous national income estimates beginning with Canada and the Soviet Union in 1925. Until 1939, nine countries had followed their ex-

4 Friedman (1940), cited from Morgan (1990, p. 127).

ample and a comprehensive League of Nations Report was published including continuous time-series for more than two dozen countries. During the war, the estimates of national income were transformed into national income accounts. This transformation made use of the fundamental idea that income and expenditure represent two sides of the same coin, i.e. that they are different ways to measure national production, and had largely been inspired by the theoretic development of macroeconomics in the 1930s (see Kendrick, 1970). As Zarnowitz (1992) points out, quarterly data for US GDP is available from the end of World War II. The development of national accounting, i.e. of a more or less direct measure of economic output, was clearly a major boost for macroeconomic forecasting. This corresponds with the fact that Morgenstern (1928, p. 30) considered the lack of such a unique measure of economic activity as a major obstacle to macroeconomic forecasting in general.[5]

We now enter the recent period of economic forecasting that is characterized by a constant proliferation of macroeconomic forecasts and a sophistication of its underlying methods. This is no surprise, since the sufficient ingredients for this boom were all present after World War II. As we have seen so far, the major forecasting methods have their roots in the first half of the 20th century or earlier: Time series methods have originated from statistics, indicator systems from the business barometers, modern econometric models from the work of Frisch and Tinbergen. In addition, the national income accounts provided continuous macroeconomic data. The development of electronic data processing boosted economic forecasting further.

The number of macroeconomic forecasts and forecasters grew steadily in the post-war decades as is nicely illustrated by figures we can take from Zarnowitz (1992). In 1959, the National Association of Business Economists (NABE) was founded with an initial membership of 322. This number rose in subsequent decades to 1682 (in 1969) and 2749 (in 1979), peaked at 3491 in 1983 and only slightly declined to 3300 in 1989. In 2004, NABE appraises it at roughly 3000 (National Association for Business Economics, 2003). However, it should be noticed that just a minority of the NABE members focuses on macroeconomic forecasting. Most business forecasters only *use* macroeconomic forecasts as inputs to establish more specific forecasts.

The rise of macroeconomic forecasting triggered the systematic collection of these forecasts and their evaluation. Yet, consistent, sufficiently large samples of forecasts are still rare as Zarnowitz (1992) notes. The first survey of macroeconomic forecasts, briefly mentioned in the previous chapter, was initiated by the

5 However, Morgenstern (1928, p. 74) also doubted that there could ever be something like a composite general index of economic activity.

journalist Joseph A. Livingston in 1946. He invited business economists to send him their forecasts of several macroeconomic variables. Livingston pursued his survey on a bi-annual basis until his death in 1992 when the survey was taken over by the Federal Reserve Bank of Philadelphia. While being the oldest survey, Livingston's compilation collects forecasts of non-professionals and the data has therefore to be judged with care (Croushore, 1997). The first survey of *professional* macroeconomic forecasters dates back to 1968: It is the ASA-NBER survey introduced in chapter 2.

## 3.2 Predictive progress of Macroeconomics

With the brief history of economic forecasting in mind, let us now turn to the more particular issue of how the predictive performance has evolved. Anticipating the general conclusion of this section we may state that the longest *comparable* time-records of macroeconomic forecasts, dating back to the 1950s, do not yield any definitive evidence for an improvement in accuracy or precision. As in the previous chapter, we will have a closer look at the different studies this conclusion is drawn from in the following.

Evaluating forecasts in a historical perspective relies on the familiar methodology of the preceding chapter: It consists in comparing error measures of appropriately classified forecasts. Before we do so, recall figure 2.1 that shows the historical record of the mean ASA-NBER real growth forecasts. This figures reminds us: Given the relatively short sample periods as well as the period dependence of forecast errors, any historical comparison has to be made and judged with care.

This said, consider some bare numbers. Zarnowitz (1992) calculated the MAE of nominal and real growth as well as inflation forecasts covering the period from 1953 to 1984 and averaged these forecast errors over a couple of (overlapping) sub-periods. The collection does not only include the ASA-NBER data but also a variety of other surveys — among others the Livingston survey, the forecasts from the Michigan research seminar on quantitative economics (RSQE) and official CEA forecasts. The conclusion drawn from this data asserts that the mean absolute errors "display no systematic upward or downward trend" (p. 523). However, when accounting for volatility by dividing the MAE by the predicted variable's standard deviation, Zarnowitz (1992) claims to find an improvement in RGNP forecasts from the late 1950s to the mid 1980s. But this improvement is solely due to a higher volatility in the more recent sub-periods as composed by Zarnowitz.[6] So, I am reluctant to infer from it an improvement of forecast accuracy

[6] I wonder why Zarnowitz did not include the late 1980s in his historical comparison, that is a period characterized by a lower volatility, although the data had been available to him.

and precision. Including the late 1980s in their evaluation but considering different sub-periods, Zarnowitz & Braun (1992, p. 36) find "no evidence here that the [ASA-NBER] forecasts on the whole either improved or deteriorated in the 1980s as compared with the 1970s."

In contrast to Zarnowitz & Braun, McNees (1992) finds that absolute forecast errors of real growth have improved in the two decades from 1970 to 1990. But he stresses that this need not be considered as an improvement of forecast performance since "much of the improvement merely reflects the fact that no turning points occurred for the 92 months between November 1982 and July 1990" (p. 29). Extending the examined period, McNees continues to investigate only the RSQE forecasts from 1953 to 1991 and finds a clear improvement if these forecasts are evaluated against a naïve benchmark. This conclusion, however, is primarily due to the bad initial performance of the RSQE models in the 1950s with RMSEs as high as an average 3.2.

Zarnowitz & Braun as well as McNees evaluated economic forecasts at the beginning of the 1990s; but what has happened since then? Is there any new evidence that might settle the question whether macroeconomic predictions have improved over time or not? Let us see what the more recent studies introduced in the last chapter have to say. First of all, Öller & Barot (2000), evaluating GDP forecasts for OECD countries from 1971 to 1998, diagnosed no significant change in accuracy and precision — neither in absolute terms nor relative to standard deviation. In fact, the RMSE/standard deviation ratio remains more or less constant at 0.88. In comparing this value to the equivalent ratio of the forecasts of the British National Institute of Economic and Social Research of 1959-67, namely 0.8, the authors convincingly demonstrate that forecast performance remained more or less the same during the last 40 years. Unlike growth forecasts, inflation forecasts have improved in terms of *absolute* errors according to Öller & Barot (2000), but they do not find any evidence for better *relative* performance. Yet this observation can be explained by the simultaneous decline in inflation in most OECD countries during the 1970s and 1980s.

Examining the forecasts of the U.S. Federal Reserve Bank from 1965 to 1989, the results of Joutz & Stekler (2000) are suited to be compared with those of Zarnowitz and McNees. Joutz & Stekler divide the 25 years under examination into five sub-periods of five years each and divide the calculated MSEs by the predicted variable's variance.[7] In contrast to the preceding studies, they evaluate forecasts with a one quarter horizon and eventually conclude that there is no

[7] They hence do not only differ from the former studies in the error measure but also in the way they account for volatility. In fact, they assume that the difficulty to forecast is not proportional to the standard deviation but to the variance of the predicted variable and — since $V = \sigma^2$ — give volatility a heavier weight.

clear evidence "that the quality of the forecasts has improved over time" (p. 23). Since Joutz & Stekler also show that the Fed's predictions are not significantly different from predictions of the ASA-NBER survey, their conclusion supports the interpretation that the thin evidence of forecast improvement Zarnowitz (1992) and McNees (1992) reported is solely due to the poor initial performance (of the RSQE models) in the early 1950s.

The historical record of German GDP forecasts is more closely investigated in two recent papers by Dicke & Glismann (2002*a*,*b*). Dicke & Glismann evaluate forecasts of the German Council of Economic Experts (GCEE) from 1964-2001 as well as the forecasts of a German business cycle institute in 1966-2001 that represents as *(melior) pars pro toto* the six leading national institutes. In contrast to the studies discussed so far, forecast quality is measured by MAPE2. According to the authors, this is justified by the facts that (i) an absolute error of 2% when growth trend equals 5% is not as severe as an error of 2% when the economy actually grows by an average of 2%, and (ii) the growth trend has changed in the period under examination. The authors show that neither the forecasts of the national institute nor the GCEE forecasts have improved in the course of the sample period. Even worse, there is some evidence that the predictive performance of the GCEE has worsened during the last four decades — a conclusion that is even better confirmed if the economy's volatility is taken into account. Unsurprisingly, this result has been challenged by the GCEE. Weidmann (2002) attempts to defend its predictions' quality mainly by questioning the forecast measure chosen by Dicke & Glismann. In addition, Weidmann uses preliminary instead of final GDP data in order to calculate the forecast errors. Using MAE and MSE and accounting for volatility, he succeeds in showing that the predictive performance of the GCEE has not changed systematically over time. However, Weidmann acknowledges that even according to his alternative measures no evidence points towards an improvement in predictive performance.

The discussion in this section has made clear that whether macroeconomic forecasts have become more accurate and precise over time is a controversial issue and a question hard to settle because of the intrinsic difficulties of any empirical forecast evaluation mentioned in the previous chapter. However, it has also become clear that there is no definite evidence of significant improvement in macroeconomic forecast performance during the last 40 years. Whether or not there was an improvement in the very first years after World War II — as the study of McNees suggests — can be left open and does not seriously spoil the following inquiry.

Let us conclude this section by listening to Lord Terence Burns, experienced forecaster and a former main economic advisor of the British government, who rounds off our discussion and links it up with the next section:

> Finally, from my interpretation of the research evidence as well as from my own investigations, the profession appears to have made very little progress in reducing the size of forecast errors over the past 30 years or so, whether for the United Kingdom, the United States, or other industrialized countries. This result is tentative because some periods are intrinsically more difficult to predict than others. It is also a surprise and a disappointment, considering the huge increase in computing capacity as well as the development of ideas in both economics and econometrics. (Burns, 2001, p. 174)

## 3.3 The use and performance of economic forecasting methods

Burns states in the above quotation that forecasting methods have become more and more sophisticated. This fact is for example illustrated by the sheer size of today's econometric models compared with those used by Tinbergen. But what role has econometric modeling played in comparison with other forecasting methods? And, secondly, how has the predictive performance of the different methods evolved over time?

The ASA-NBER data contains, as mentioned earlier, information about the predictive methods the surveyed forecasters applied. This enables Zarnowitz (1992) to assess the relative importance of alternative methods. His clearest result states that there is no single stand-alone method, i.e. different methods are always used together in order to complement each other. The most frequently used method is the informal approach involving a large portion of judgment. More than 70% of the forecasters used this method and more than half of them ranked it first in all three examined time-periods (see p. 38 above). The second most frequently used method in the late 1960s was the leading indicator method. But the use of this approach declined sharply in the subsequent years. Leading indicators have always predominantly been used as a second or lower ranked method. The same holds for anticipation surveys while here the decline is even sharper.

Roughly half of the forecasters used econometric models provided by other institutions and about a quarter based their predictions on their own econometric models, both ratios following a slight upward trend. In the same time, econometric modeling has been judged more and more important for forecast production. However, this trend is more apparent in the first half of the 1970s than in the second. This fact might be due to the severe critique of econometric forecasting issued by Lucas (1976) which will be discussed in chapter 9.

In his study discussed in the preceding chapter, Armstrong (1984) came to the result that more complex econometric models do in general not yield predictions of higher quality. This conclusion can be interpreted — and so it is by Armstrong himself — as showing that progress in model building does not necessarily bring about progress in predictive performance.

Unfortunately, Zarnowitz (1992) does not analyze the historical record of the different methods although the ASA-NBER data would allow for such an examination. We may nevertheless argue — on the basis of the established facts, namely that (i) there is no significant difference among the methods with respect to their performance, (ii) there is no difference among the forecasters with respect to their performance and (iii) there is no improvement in overall forecasting accuracy — in favor of the conclusion that the forecasting performance of the individual methods has not improved significantly over time. Had it improved significantly, the facts (i)-(iii) could hardly be explained.

## Main observations

**Observation 8 (Forecasting boom)** *The forecasting literature and industry has constantly been growing since the end of World War II.*

**Observation 9 (Lack of improvement)** *The quality of macroeconomic forecasts has not improved during the post-war period, in any case not during the last 40 years.*

**Observation 10 (Role of econometrics)** *Although not being a predominant method, econometric models have become more and more important for the production of macroeconomic forecasts.*

**Observation 11 (Methods' performance)** *It may be inferred from the evidence that there has been no significant improvement in any macroeconomic forecasting method either.*

# Chapter 4

# Forecast performance in natural sciences


## Summary

This chapter evaluates the predictive performance of thermodynamics in distinguishing between experimental and non-experimental contexts. Two historic case studies, namely on the latent heat of fusion and the ideal gas law, warrant the conclusion that experimental thermodynamics is characterized by predictive progress *in depth* as well as *in scope*. The non-experimental case is explored by examining attempts to predict the climate. Climatology's predictive performance is found to be poor.


## 4.1 Progress on the experimental level: Specifying latent heat of fusion

It is an undisputed common-place that (i) natural sciences provide *some* predictions of utmost accuracy and that (ii) there has been *some* tremendous progress in the predictive performance of natural sciences. This chapter is therefore not meant to either prove or disprove these obvious facts. Instead, it aims at specifying the reasons as well as the limits of natural sciences' predictive success.

A paradigm underlying natural sciences is that natural phenomena can be studied under artificially produced, well controlled situations in shielded environments. This is the essence of the experimental method. However, natural sciences' domain of inquiry is by no means restricted to such experimental situations. Accordingly, this and the following section focus on the predictive performance and progress under experimental conditions while the third section is devoted to forecasting in natural sciences without experimental control.

Before embarking upon the empirical assessment, we should notice that there is a close conceptual link between the method of experimentation and predictive performance. A central methodological imperative of experimental science states that experiments must be replicable. But this alone already implies conditional predictability! For predictions of the outcome of experiments hold conditional to the experimental setting and some specific interventions. If these conditional forecasts fail for some experiment, it will not be replicable and, as a consequence, will not be considered as a valid experiment at all. Consequently, experimental success implies predictive success, and progress on the experimental level is a sufficient condition for predictive progress. Yet, it does not seem to be a necessary condition, and it is thus an open question to what extent predictive progress in natural sciences is actually due to improved experimental skills.

This question shall now be explored in a historical case study on the measurement of latent heat.[1] The Scottish physicist Joseph Black (1728-99) discovered, while deliberating on experiments by Fahrenheit on phase transitions in the middle of the eighteenth century, the phenomenon of latent heat, namely that a certain quantity of heat is required to melt a solid (latent heat of fusion) or to vaporize a liquid (latent heat of vaporization). Without clear concepts of heat and energy, this discovery could then be stated in the following way: A certain mass $m$ of ice at a temperature of 0°C is completely transformed into water of 0°C if water of mass $m$ and temperature $T_F$ is added to the ice. As soon as $T_F$ is numerically determined, this becomes a conditional prediction, including a point forecast of the water's temperature.[2]

Black attempted to measure the latent heat of fusion of water but never published his results so that the credit for the calculation of the latent heat of water is usually attributed to Karl Wilcke (1732-96) who, in 1772, independently determined the value to be 72°C; actually, Black found it to be 140°C Fahrenheit, i.e. $77.\overline{7}$°C (see Cardwell, 1971). The modern value is 79.7 cal/g, but we should notice that it is of no importance for the predictive performance whether the obtained values are close to the today agreed upon value or not. What *is* of importance is that the experimental result is replicable. For an experimental setting that consistently yields exactly the same, nonetheless 'completely wrong' value (relative to what we currently believe to be the correct value) allows for much more accurate and precise prediction than a setting that gives rise to highly varying values although its mean outcome equals the 'right' value.

In order to determine the numerical value of latent heat, Wilcke mixed equal

[1] This as well as the following case study is primarily based on articles of the excellent *Dictionary of Scientific Bibliography* Gillispie (1971) which are not explicitly cited.

[2] $T_F$ is not exactly the latent heat of water which is nowadays defined as the ratio of energy required to melt a certain mass of ice, and is measured in cal/g. Yet, as (i) the energy required to raise the temperature of water by one degree is approximately constant and independent of its absolute temperature, and (ii) as 1 cal denotes the energy needed to heat 1 g of water from 0 to 1°C, the absolute value of $T_F$ approximates the latent heat of fusion as measured in cal/g.

masses of melting snow (0°C) and water at temperature $T$. Thus, if no heat were needed to melt the ice, the expected temperature of the mixture would be the mean of the initial temperatures,

$$T_{\text{mix}} = \frac{T+0}{2} = \frac{T}{2}.$$

The observed temperature $\theta$, however, varied from this expected value. In addition, Wilcke noticed that the loss of temperature,

$$M = T_{\text{mix}} - \theta \tag{4.1}$$

remained constant for different initial temperatures $T$. He therefore assumed that some heat was used to melt the ice so that the hot water's temperature was decreased by some amount, $T_F$, without increasing the cold (melted) water's temperature. The effective temperature that heats the cold (melted) water thus is $T - T_F$ and the observed equilibrium temperature equals,

$$\theta = \frac{T - T_F}{2}. \tag{4.2}$$

Simple transformations and a substitution of (4.1) into (4.2) yield,

$$\begin{aligned} T_F &= T - 2\theta \\ &= 2T_{\text{mix}} - 2\theta \\ &= 2M. \end{aligned}$$

Wilcke obtained the empirical value of $M = 36\frac{3}{28}$ by experimentation and derived that the latent heat of one unit mass ice is 72°C. This finally enabled him to make specific conditional forecasts for similar experimental settings.

Wilcke's method of mixing was commonly used during the whole first half of the nineteenth century. It had been improved by using better thermometers and more accurate balances so that its precision increased constantly. In 1843, the French thermo-chemists Frédéric Hervé de La Provostaye and Paul Desains obtained a value of 79,2 cal/g by the method of mixing. Four years later, Germain Henri Hess (1802-1850) performed a series of experiments using the same method and obtained, according to Médard & Tachoire (1994, p. 383), a value of 80,34 cal/g. The increasing number of decimals indicates that the measurements have become more and more precise. Consequently, the outcome of such mixing experiments could be forecasted with ever higher precision and accuracy.

Predictive progress in experimental situations is not simply a linear process driven by more and more accurate instruments. In fact, a lot of different experi-

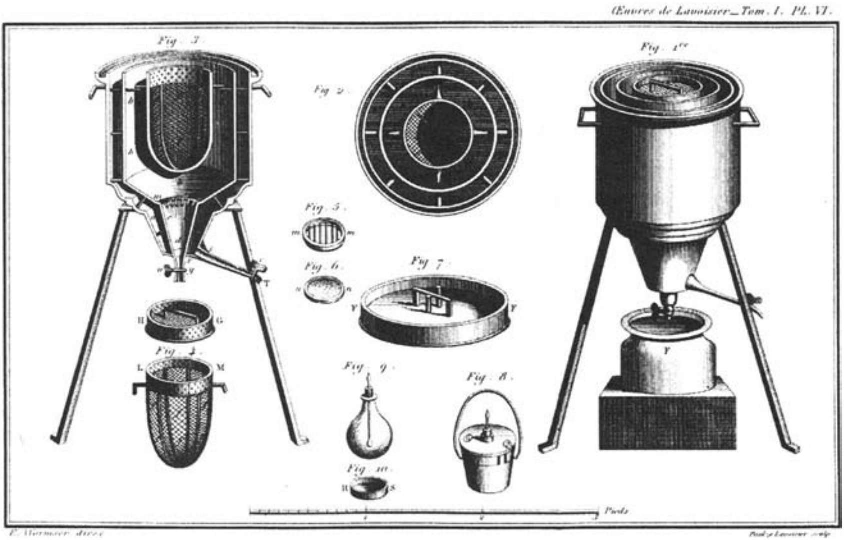


Figure 4.1: The ice-calorimeter of Lavoisier and Laplace. The volumes $a$ and $b$ contain the ice, and the probe is inserted in $f$. Only the ice in $b$ is melted by the probe and its melting water is collected in a case situated directly under the instrument. The outer ice-shield $a$ has only one purpose: To isolate the inner ice from the surrounding environment and to ensure that the ice in $b$ is melted by the energy released from the probe only. (Source: Lavoisier, 1789)

mental settings and instruments may compete with each other. The ice-calorimeter developed as early as in 1780 by Antoine Lavoisier in collaboration with Pierre Laplace was such a competitor to the mixing method. This instrument was designed to determine the specific heat of substances and its corresponding measurement theory involved the concept of latent heat of fusion: The probe is heated and placed in an ice cube until it completely cools down to 0°C. The heat that has been emitted from the body is then determined by weighing the amount of ice that has melted (see figures 4.1 and 4.2). Lavoisier and Laplace saw that in order to determine the heat emitted by the body numerically, they needed a 'numerical fixpoint',

> La quantité de calorique nécessaire pour fondre une livre de glace nous a fourni cette unité: or, pour fondre une livre de glace, il faut une livre d'eau élevée à 60 degrés du thermomètre à mercure divisé en 80 parties, de la glace à l'eau bouillante; la quantité de calorique qu'exprime notre unité est donc celle nécéssaire pour élever l'eau de zéro à 60 degrés. (Lavoisier, 1789, p. 291)

The authors expressed the latent heat of fusion in terms of the scale introduced by René Antoine Réamur (1683-1757) around 1730, their value equals

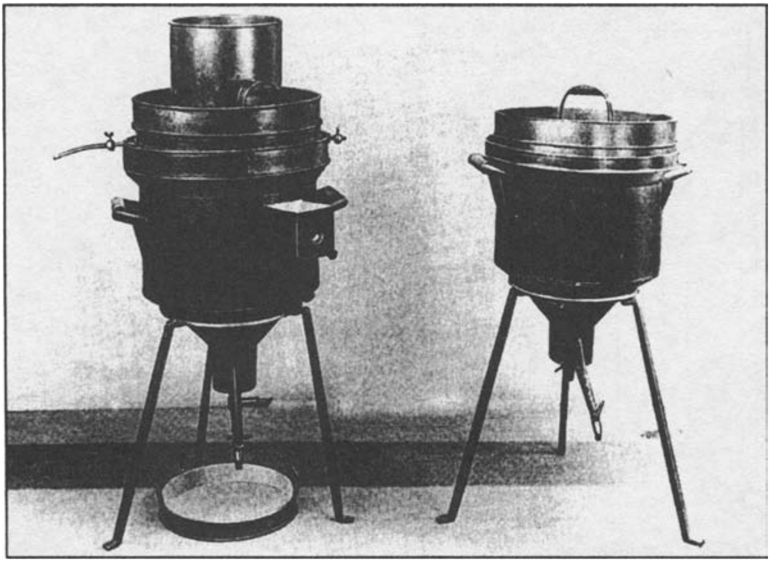


Figure 4.2: The original instrument of Lavoisier and Laplace that can be viewed in the Musée des Arts et Métiers at Paris.

75°C. However, they do not specify how they obtained this value. Did they use the ice-calorimeter itself? All we may say is that this would have been possible. Lavoisier and Laplace themselves explain how to measure the heat emitted by a liquid,

> Quant aux fluides, on les renferme dans des vases de matière quelconque, dont on a préalablement déterminé la chaleur spécifique: on opère ensuite de la même manière que pour des solides, en observant seulement de déduire, de la quantité totale d'eau qui a coulé, celle due au refroidissement du vase qui contenait le fluide. (p. 290)

If this liquid would have been water at temperature $T$ and of mass $m$, the latent heat could have been calculated by $T_F = \frac{m}{m_{\text{ice}}}T$ where $m_{\text{ice}}$ is the (adjusted) mass of the melted ice.

One century later, in 1870, Robert Wilhelm von Bunsen reverted to the idea of Lavoisier and Laplace and developed an ice-calorimeter that allows to measure the specific heat with utmost accuracy (Bunsen, 1870). Figure 4.3 is a schematic drawing of his instrument. In contrast to Lavoisier and Laplace, Bunsen did not determine the quantity of heat emitted from a body when cooled to 0°C by weighing the corresponding amount of melted ice but by measuring the volume change of the ice-water mixture that surrounds the body. For if a probe is inserted into the instrument (cylinder $a$), a certain amount of ice melts, the volume contracts

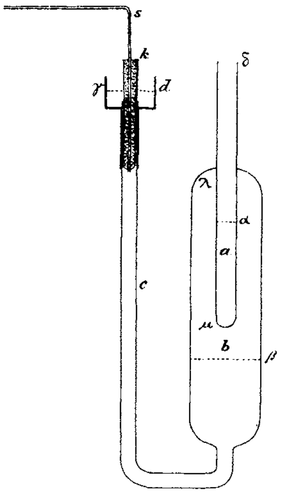



Figure 4.3: The ice-calorimeter of Bunsen. The glass cylinders $a$ and $b$ are filled with distilled water between $\alpha$ and $\mu$ and between $\lambda$ and $\beta$ respectively. The remaining volume of $b$ including the tube $c$ and the scale $s$ is filled with mercury. As final preparatory steps, the water in $b$ is frozen at 0°C and the whole instrument is embedded in a sufficiently big box filled with snow. (Source: Bunsen, 1870)

and the mercury in the scale descends. The latent heat of fusion of water can be calculated in three steps: (1) Measure the volumetric change that occurs after hot water has been introduced in the cylinder. (2) Calculate the mass of the melted ice. (3) Calculate the latent heat as indicated above for Lavoisier's calorimeter. This procedure (in step 2) requires knowledge of the specific weights of water and ice. Bunsen performed experiments on his own to determine the specific weight of ice. He was subsequently able to find the latent heat of fusion to be equal to 80,025 cal/g which is in fact the mean of two experiments that yielded 80,01 and 80,04 respectively. Bunsen has thus constructed an experimental setting that allowed him to predict the outcomes with more accuracy than the old ice-calorimeter.

As already mentioned above, predictive progress is not linear. New instruments sometimes do not allow for as accurate predictions as well-established ones but nevertheless bear a large potential which might unfold over time. This is the case for the calorimeter invented by the physicist Charles Person (1801-84) in 1850 (see Médard & Tachoire, 1994, p. 383). Measuring the latent heat of fusion, he obtained a numerical value of 80 cal/g. This was less precise than the available values, but he used an innovative method. The main idea of Person's so-called adiabatic calorimeter is to heat or cool the outer wall of the instrument so that there is no heat flow between the instrument and its environment. The quantity of

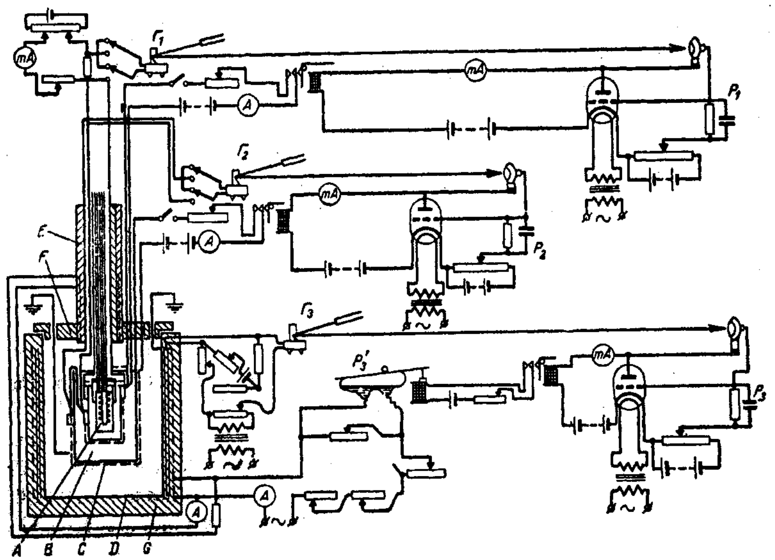



Figure 4.4: Blueprint of a modern adiabatic calorimeter with an electronic control system that ensures absence of net heat flows: (A) is the calorimetric vessel, (B) and (C) are shields, (F) and (G) heaters. (Source: Kagan, 1984)

heat that is exchanged between the calorimeter and a certain probe is then inferred from the quantities of heat involved in the heating and cooling process. With the whole calorimeter put in a reservoir filled with hot or cold water, Person heated and cooled his instrument by raising or lowering it according to whether the temperature of the calorimeter's outer wall is lower or higher than the reservoir's. Via the use of a differential thermometer, the experimenter made sure that the instrument's temperature equals the environment's temperature. Person's ideas were rediscovered in the late 19th century and are still used in modern adiabatic calorimeters such as presented in figure 4.4.

A further rediscovery relates, according to Médard & Tachoire (1994), to an old idea of James P. Joule that was used to improve calorimeters in the beginning of 20th century. Joule's original idea was this: The thermo-physical properties of an object can be studied by (i) connecting the object with a second object whose thermo-physical properties are well-known and (ii) measuring the heat flows between the two objects. So-called differential calorimeters implement this idea. The development of thermoelectric elements allowed to measure temperature with utmost accuracy and to produce micro-instruments that could be employed in extreme situations such as high pressures. Initiated by Albert Tian (1880-1972) and Edouard Calvet (1895-1966), this led to the development of modern high-

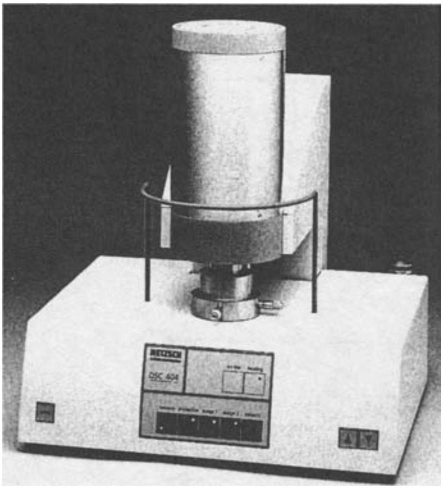


Figure 4.5: A high precision DS-calorimeter produced by *Netzsch*. (Source: Netzsch GmbH)

precision instruments: Modern differential scanning calorimeters (DSC) heat two tangent probes independently of each other so that the temperature is equal and in fact no net heat flow occurs (Richardson, 1984). The difference in energy added to the two probes allows to deduce the thermo-physical properties of one probe given those of the other. DSCs can produce huge quantities of data with ease that could not possibly have been obtained by adiabatic or drop calorimetry[3] alone. Modern calorimeters like those in figure 4.5 allow for forecasts with a precision that is worlds apart from the performance of 18th or 19th century experimental physics. In the public database of the US National Institute of Standards and Technology, the heat of fusion of water — to be precise: the standard enthalpy of fusion — is given with an accuracy of four decimal places.

The history of latent heat makes clear that there is a progress in forecast performance due to the improvement of experimental skills and techniques. But is the diagnosed progress solely due to improved experimental skills? The answer is yes, as a simple argument shows: Assume, for the sake of argument, that there were some predictive progress that did not result from improved experimental techniques. Accordingly, we should be able to predict the outcome of Wilcke's or Bunsen's experiments with more accuracy and precision than they did. But that is not the case. Wilcke's mixing experiment does not allow for more accurate predictions today than it did 200 years ago! Hence, the predictive progress

[3] Which is in fact the modern variant of Wilcke's good old mixture method, see Ditmars (1984) for more.

*only* consists in devising more precise instruments and more sophisticated, more effectively shielded experimental settings.[4]

## 4.2 Progress on the theoretical level: The ideal gas law

Hitherto, we have studied how forecasts of one and the same phenomenon, namely the melting of ice, improved over time due to the invention of techniques which allowed for ever more precise replication of that phenomenon. Successful prediction solely stemmed from successful replication. This kind of predictive progress insofar as restricted to one particular phenomenon might be labeled progress *in depth*. It does not involve any predictive model and is certainly not the only kind of progress we would expect to observe in natural sciences.

In fact, we would expect that the sciences enable us to forecast more and more different phenomena (For instance: not only to forecast the equilibrium temperature of a mixture of equal masses of ice (at 0°C) and water (at $T_F$) but of mixtures of arbitrary masses, arbitrary temperatures, different substances, ...). The following case study on the ideal gas law suggests that such predictive progress *in scope* which seems to require improved predictive methods and thus a theoretical advancement, ultimately depends on and is limited by our experimental skills, too. Advancement on the theoretical level which consists in the formulation of general laws that may be used to derive (more) successful predictive models for (more) different phenomena does not entail that the range of successfully predictable empirical phenomena widens.

The British physicist and cofounder of the Royal Society, Robert Boyle (1627-91), discovered, while experimenting with air and vacuum in the late 1650s, that air is permanently elastic. This statement had been asserted more precisely in a later edition of his *New Experiments* in the version which subsequently became known as "Boyle's law". It states that the pressure $P$ of a gas is inversely proportional to its volume $V$: $PV = \text{const}$. Boyle's law was discovered for a second time by the French scientist Edme Mariotte who published the results of a series of experiments in 1679. Mariotte used a mercury barometer to establish that the ratio of the volumes of two gases is equal to the ratio of weights that exert pressure upon them — a statement equivalent to Boyle's law. Although Mariotte treats this relationship as if it were well-known, there is no reference to Boyle, and in France Boyle's law is usually referred to as Mariotte's law.[5]

4 I concede that the development of new instruments typically arises out of a better general understanding of physical processes and that theoretical advancement in natural sciences might thus indirectly contribute to predictive progress. Still, the primary cause is the improvement of technical skills.

5 By the way, Boyle's law was, according to Cardwell (1971, p. 4), first discovered by

While Boyle and Mariotte investigated the relationship between pressure and volume of air, there has been little research on the role of temperature in this context during the 17th and 18th century. Guillaume Amontons (1663-1705) seems to have been the first to measure the thermal expansion of elastic fluids. Without any thermometer or an established thermometric scale he estimated that air expands by one third between 'cold' water and its boiling point (see Cardwell, 1971, p. 19). Yet there was still no agreement on the expansion coefficient of air at the end of the 18th century (Cardwell, 1971, p. 122).

Thus, the scene was set for the French chemist Joseph-Louis Gay-Lussac (1778-1850). He devoted his first investigations in 1801-02 to the thermal expansion of gases and was able to conclude that "equal volumes of all gases expand equally with the same increase in temperature" (Cardwell, 1971, p. 126) thus stating what we call today "Gay-Lussac law".[6] Or formally: $\frac{V}{T} = const.$ In addition, Gay-Lussac calculated that air when heated from 0°C to 100°C expands by a volume equal to 3/800 of the volume at 0°C. Similar experimental research was carried out by the English physicist and chemist John Dalton in 1802, who found the coefficient to be equal to 1/266. However, Dalton did not postulate a linear expansion between the two temperature points of his observation and consequently did not accept Gay-Lussac's law.

Since Boyle's law holds under condition of constant temperature and Gay-Lussac's law requires constant pressure, there is a single unifying law of which both become two special cases. This single law is the ideal gas law, namely

$$pV = RT \tag{4.3}$$

where $R$ is a universal constant. This equation was a concise, summarizing statement of the experimentally observed regularities obtained by Boyle, Gay-Lussac and others. The empirical data could be fitted into this equation. In the same time, the universal law allowed to derive infinitely many predictions for yet unobserved phenomena. Still, it is another question whether it widened the scope of successfully predictable phenomena. For not only was it an open question whether the ideal gas law yields correct predictions for, say, extreme pressures, but improved experimental skills were also necessary to fabricate such conditions.

The French chemist Pierre Louis Dulong (1785-1838) was working on heat throughout his career. In a paper in 1815, Dulong and his colleague Aléxis T.

Richard Towneley and Boyle himself called this law 'Towneley's hypothesis'.

6 As in the case of Boyle's law, it is not clear who really discovered Gay-Lussac's law for the first time: The French experimental physicist Jacques-Alexandre-César Charles seems to have found the Gay-Lussac law already in 1787, but he neither published his results nor calculated the expansion coefficient. In addition, his experimental procedure was significantly improved by Guy-Lussac.

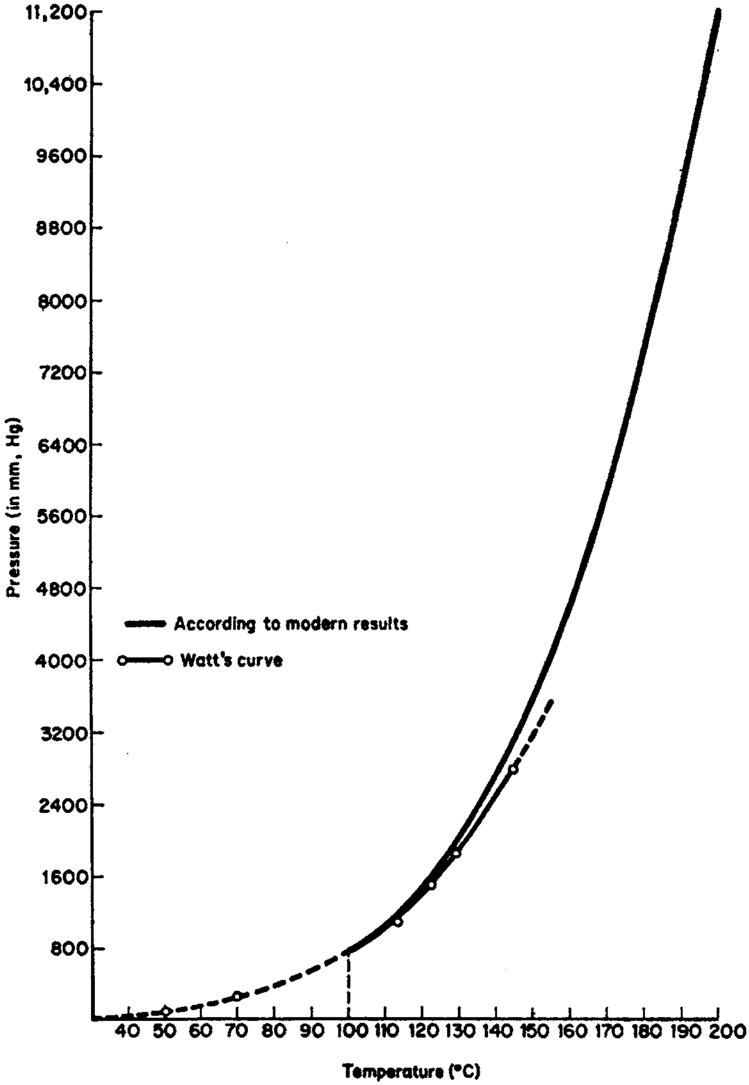



Figure 4.6: Variations of steam pressure with temperature. (Source: Cardwell, 1971)

Petit (1791-1821) confirmed the Gay-Lussac law for a temperature range between -40°C and 200°C (see Cardwell, 1971, p. 137). Fifteen years later, Dulong worked with the physicist Dominique Francois Jean Arago (1786-1853) in a commission charged with studying the pressure of steam under high temperatures. This research was prompted by the French government's concerns relating to the security of boilers in steam engines. In the course of their study they proved the validity of Boyle's law for high temperatures and pressures up to 27 atmospheres. Research on the behavior of steam under high pressures had been conducted half a century before Arago and Dulong's research by the inventor of the steam engine himself, James Watt (1736-1819). Watt knew that water might boil under low pressure at temperatures less than 100°C and aimed at determining the pressure of steam generated at 50°C or 70°C. He performed several experiments at temperatures above 100°C and then extrapolated his results to lower degrees. His findings are shown in figure 4.6 which nicely illustrates — by the comparison with modern data — how the range of predictable phenomena has indeed increased.

In light of its application to steam, it has always been clear that the ideal gas law is not empirically adequate for all temperatures. A clear limit of scope is reached when a gas is liquefied: The relationship breaks down in phase transitions. It is therefore no surprise that several experimenters in the first half of the 19th century found more or less conclusive evidence on different deviations from the ideal gas law. Among them were the Danish physicist Hans Christian Oersted, the German Gustav Magnus (1802-1870) and Johann Natterer (see also Darmstaedter, 1908). However, the most important experimental physicist and chemist in this context is Henri Victor Regnault (1810-1878) who, in his studio in Paris, produced data of unseen quality and quantity. After a series of experiments, he finally came to the conclusion in 1853 that the coefficient of thermal expansion is not a constant for all gases but varies for different gases in contradiction to the ideal gas law. In addition, he showed that there are small empirical deviations from the values predicted by the ideal gas law and that these deviations become the more important, the more the gas approaches its temperature of condensation.[7] Regnault's data sets were of utmost value for the scientists in the second half of the 18th century and were used in major contributions to thermodynamics such as by Clausius (1850) or van der Waals (1988).

The experiments of the Irish chemist Thomas Andrews (1813-1885) and the French physicist and engineer Louis Pierre Cailletet (1832-1913) in the liquefaction of gases and their discovery of a 'critical point' at which the substance enters an intermediate state between fluid and liquid had two important consequences. First of all, they invalidated the idea that the ideal gas law is at least very accurate

[7] Clausius (1850, p. 378) reports these results.

for *some* (so-called permanent) gases. Secondly, the question arose whether there is a more fundamental equation of state that not only covers the gas phase but the liquid (and solid) phase as well.

On the theoretical background of Clausius' molecular-kinetic deduction of the ideal gas law (which we will reconstruct in chapter 10) on the one hand and the experimental background just described on the other hand, the Dutch Johannes Diderik van der Waals (1837-1932), in his dissertation originally published in 1873 (van der Waals, 1988), developed a new equation of state establishing a relationship between pressure, volume and temperature of a substance,

$$(P + \frac{a}{V^2})(V - b) = RT \tag{4.4}$$

where $a$ and $b$ are universal constants that can be estimated by evaluating the behavior of gases. His equation of state allowed to explain the 'critical temperature' discovered earlier by Andrews and corrects some deviations of the ideal gas law. Yet, it was clear from the beginning that the so-called van der Waals equation is only an approximation, too. We should notice that van der Waals' corrections can be neglected for (i) high volumes or (ii) high temperatures and low pressures respectively, so that the ideal gas law (4.3) is a limiting case of (4.4).

So far, the gas laws had only been tested by either holding temperature or pressure constant. Andrews was the first who varied temperature when examining the compressibility and expansion of gases. Yet the coefficients of compressibility — $Z = pV/RT$ — and of expansion when both pressure *and* temperature varied were not known. It had been exactly this experimental enterprise the French physicist Emile Amagat (1841-1915) embarked upon in the late 1870s. Using glass tubes, he was able to experiment at pressures up to 400 atmospheres. The results were then used to build a manometer with free-moving pistons in viscous liquids that allowed him to experiment at pressures of 3,000 atmospheres. These experiments were performed in a mine near St. Etienne, 400m under ground. This was a tremendous extension of the available studies by Regnault (up to 30 atmospheres) and Andrew (up to 110 atmospheres). Amagat therefore gave an important empirical input to theory development and testing such as carried out by van der Waals.

I mentioned above that the van der Waals equation, too, had difficulties to precisely predict the outcomes of experiments. It follows for instance from the fully specified equation that the fraction $\frac{RT_c}{p_c V_c}$ is, at the critical point, a constant for all gases, equal to 2.67. However, Sydney Young (1857-1957), English physical chemist, found the actual value to be ranging between 3 and 4. Such deviations gave rise to further modifications of the equation of state which were again sub-

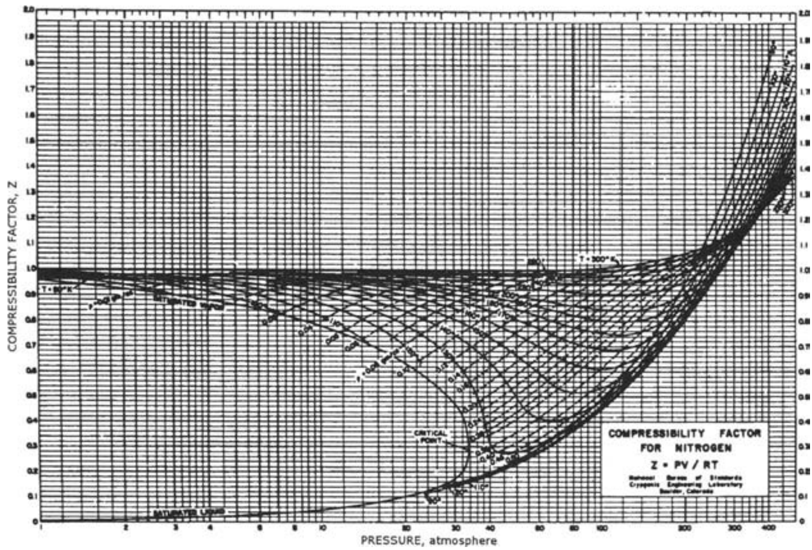



Figure 4.7: Isotherms in a Z-p diagram of nitrogen. (Source: Sonntag & Wylen, 1991)

mitted to empirical tests ...

Thanks the modern instruments presented in the last section (figure 4.5), we today have an enormous amount of data describing the $p$-$V$-$T$ behavior of thousands of substances. Figure 4.7 is an isothermal plot of the compressibility factor $Z$ and illustrates the deviation of a real gas (nitrogen) from the ideal gas law: According to the ideal gas law the plot should be a flat line. Another way to visualize the $p$-$V$-$T$ data is to plot it in a 3-dimensional diagram. Figure 4.8 shows a schematic and an approximated diagram of the $p$-$V$-$T$ surface of water.

Given the complexity of the thermodynamic behavior of substances illustrated by these diagrams, it might not be surprising that there is no single equation of state that successfully describes the behavior of gases and liquids. On the contrary, there is a bulk of different equations of state, each of them fitting more or less to some substances. Sonntag & Wylen (1991) distinguish three types of such equations: general equations of state in the tradition of van der Waals, empirical equations of state using many constants which have to be determined empirically[8], and theoretical equations of state that are derived from statistical thermodynamic

[8] One of the best known empirical equations of state is the Benedict-Webb-Rubin equation with eight empirical constants:

$$p = \frac{RT}{V} + \frac{RTB_0 - A_0 - C_0/T^2}{V^2} + \frac{RTb - a}{V^3} + \frac{a\alpha}{V^6} + \frac{c}{V^3T^2}(1 + \frac{\gamma}{V^2})e^{-\gamma/V^2} \quad (4.5)$$

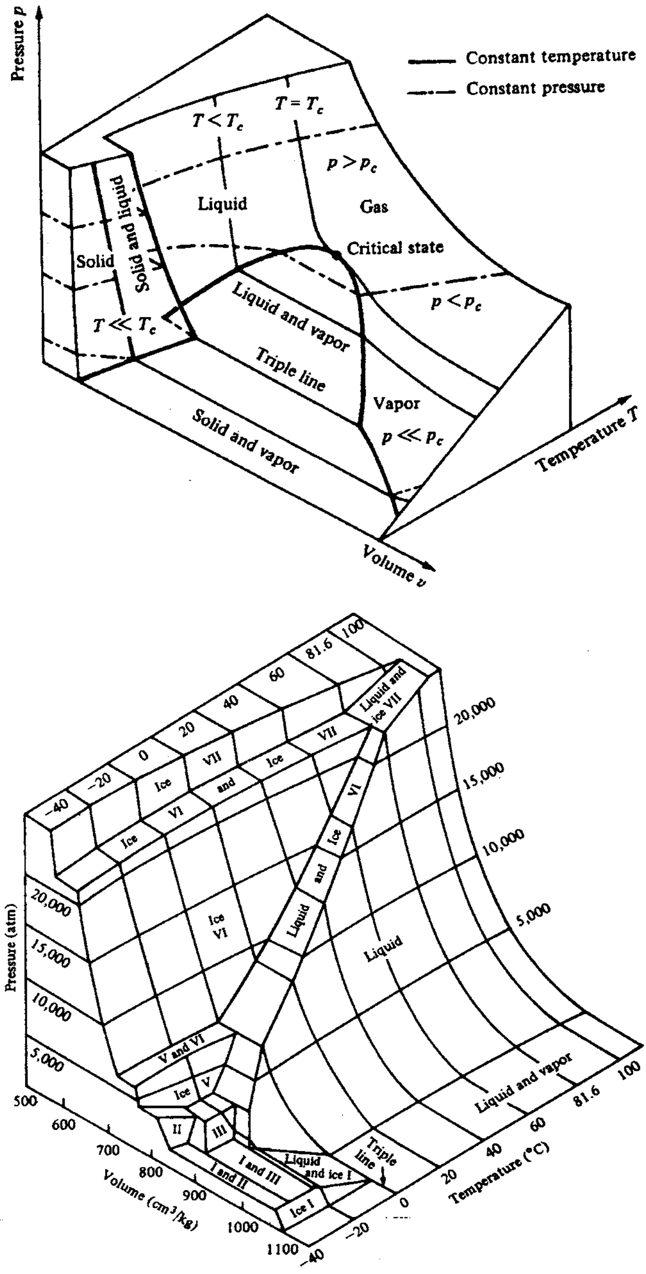



Figure 4.8: The *p-V-T* surface of water in a schematic and an approximated plot. (Source: Gyftopoulos, 1991)

theory. Such equations serve the only purpose to 'compress' the huge amount of empirical data and make it available in a convenient form. This is especially true for the purpose of programming. However, if utmost accuracy is required, the best an engineer can do is to look up the exact value in a table of thermodynamic data. And Baehr (2000) predicts that such thermodynamic tables will still be very useful in the future.

Our historical investigation has shown that there is not only predictive progress in depth (accuracy and precision) but also in scope. The range of phenomena that have been predictable became wider and wider — starting with the compression of air under several bar and leading to the complex $p$-$V$-$T$ behavior of thousands of substances under extreme situations. However, this progress crucially depended on the experimental techniques available. The scope of successfully predictable phenomena is bounded by the capacity to technically fabricate and control the phenomena. If Boyle had predicted the volume-reduction of air compressed by 3000 atmospheres, that forecast would have been anything but credible. It were Regnault's carefulness, Andrew's accuracy and Amagat's creativity that widened the boundary of successfully predictable phenomena and not Boyle's, Gay-Lussac's or van der Waals' postulation of a new, general gas law.

General laws are very useful and probably indispensable insofar as they enable us to summarize the unmanageable amount of experimental regularities we can produce. Predictive models derived from such a general law are successful to the extent that the general law accurately captures the observed and fabricated phenomena. Given this functional characterization of scientific theories, they obviously do not contribute to predictive progress in scope.[9]

## 4.3 Progress outside of laboratories: Climate predictions

*The Science of Climate*

We have seen above that experimental success necessarily implies predictive success and that, as a matter of fact, experimental progress seems to be largely responsible for the predictive progress in natural sciences. Yet, it would be too strong a conclusion to state that an improvement in experimental skills is necessarily re-

[9] Some philosophers might tend to object that this is too narrow a functional characterization of scientific theories. Pinpointing the theory-ladenness of observation, they might argue that theories not only summarize regularities between but already shape the phenomena. And I agree that theories might turn out to be crucial from this enlarged perspective. But what we are then talking about are not merely theories in the sense of formulas that enumerate general laws but rather whole conceptual frameworks. Conceptual invention might, as the example below will show, improve our predictive capacities. Yet this just repeats what I have stressed in the introduction: That all limits of forecasting and their explanations hold conditional to a linguistic framework.

quired for an improvement in predictive success, for counter-examples are easily found. It was for example not a technical but a mere conceptual invention which improved the forecasting performance of ancient astronomers when they started distinguishing between stars and planets. It is therefore an empirical question to which extent science can be predictively successful under non-experimental conditions and we are now going to extend our thermodynamic case studies to climatology in order to cover such cases.

In fact, there has been a fruitful relation between thermodynamics and climatology from the very beginning of these sciences. The English astronomer and geophysicist Edmund Halley (1656-1742), well-known because the comet he observed in order to test Kepler's law is named after him, explained the occurrence of trade winds by making use of the phenomenological law that hot air rises and cool air descends. And this before such fundamental concepts as latent heat or the ideal gas law were well-established. In the 18th century, Erasmus Darwin (1731-1802) made use of the advanced thermodynamic knowledge to explain several other meteorological phenomena. The explanation of the co-existence of sharply contrasting climates on a small area, such as the snow-covered summits of the equatorial Andes and the tropical forests at their foothills, involved for instance the ideal gas law: Cold air is compressed when falling from the summits and hence heated, while, inversely, warm air expands when it rises and according to the gas law cools down. Darwin exploited this idea furthermore in trying to relate warm and cold weather with atmospheric pressure.[10] It is not surprising that advances in thermodynamics had positive effects on meteorology and climatology for these sciences can be considered as an application of thermodynamic theory to one particular system: the earth.

Our investigation into the predictive performance of thermodynamics under the absence of experimental control will focus on climatology. The task to predict the future climate has become urgent with the recognition that collective human behavior can have a significant impact on our climate system: such as for instance causing the depletion of the ozone layer. A human-induced increase of the greenhouse effect triggered by the emission of greenhouse gases (GHG) might, however, be even more grave than that example. In 1988, the World Meteorological Organization and the United Nations Environment Programme jointly established the Intergovernmental Panel on Climate Change (IPCC) as a common global roof for climate sciences. The IPCC is charged with collecting and assessing scientific information on climate change. It is also providing scientific advice to the Conference of the Parties to the UN Framework Convention on Climate Change. In 2001, the IPCC has published its Third Assessment Report (TAR) that gives a

---

[10] For more details on the impacts of thermodynamics on meteorology see Cardwell (1971).

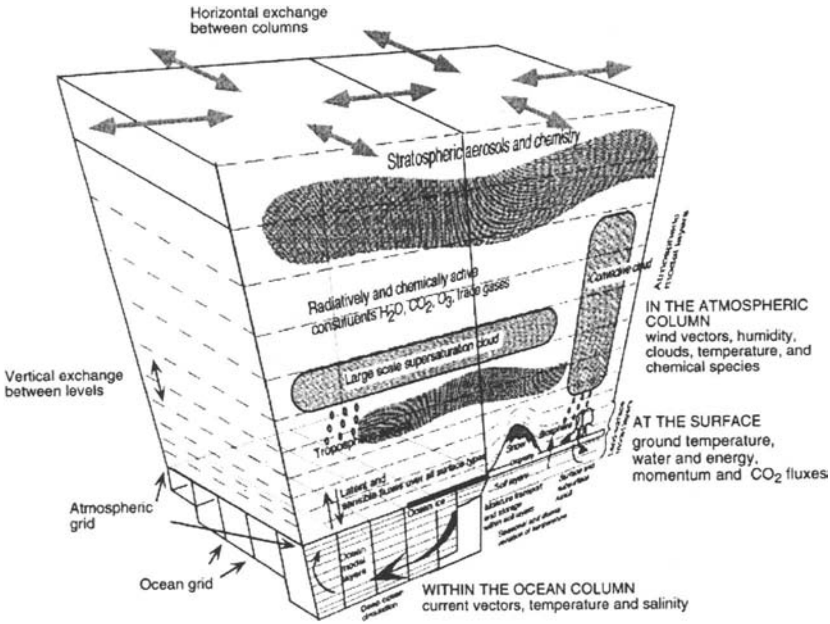



Figure 4.9: Illustration of the basic design of a general circulation model. (Source: McGuffie & Henderson-Sellers, 2001)

comprehensive overview of the state of our scientific knowledge regarding our climate. The contribution *Climate Change: The Scientific Basis* of Working Group I to this report will serve as a basis for our assessment of the predictive performance of climatology.[11]

Before assessing the predictive performance of climatology, let us briefly consider how these predictions are produced.[12] A comment on figure 4.9 which illustrates the overall design of a modern global climate model will serve this purpose. Such a model consists of several main components that represent (i) the atmosphere, (ii) the ocean and (iii) the land-surface. These components are independent models on their own that are coupled by defining energy and mass exchanges such as evaporation, rain, radiation, etc. in order to yield a comprehensive general circulation model (GCM). Yet, due to the complexity of the submodels, most complex GCMs consist of coupled atmosphere- and ocean-models only (AOGCMs). Land-surface models are not dynamically integrated into these models; instead, land-surface data enters into AOGCMs as a boundary condition.

[11] I mark citations from the Third Assessment Report throughout this inquiry both by citing the contributing authors or by just using the abbreviation "TAR".

[12] McGuffie & Henderson-Sellers (2001) is a good introduction to climate modeling.

Each sub-model is characterized by a structure similar to the first numerical models constructed for weather prediction. The climate component (atmosphere, ocean, land-surface) is divided into cubic subspaces by a three-dimensional grid. Characteristic variables as temperature, wind- or current-vectors, humidity etc. are attributed to each subspace. Finally, horizontal and vertical interdependencies between the subspaces are defined. Now, for given initial conditions, the model can be solved by a time-step procedure similar to the numerical solution of econometric models we discussed in chapter 2.

Still, this first sketch is too simple a picture insofar as it suggests that climate models can be *deduced* from thermodynamics and chemistry which is not the case. Or, the other way around, they cannot be reduced to these sciences. The sheer number of climate relevant processes necessitates simplification: A model that includes all relevant components and aspects could not even be solved on the fastest super-computers. The necessity for leaving out certain aspects of the climate system implies that choices have to be made in the course of the model construction. Additional options emerge because other physical processes that are not represented in some model (for instance because they emerge on sub-scale levels) may be integrated through additional equations that are added to the grid-structure of the model. The definition of the concrete mathematical and numerical representation of these processes is called parametrization and leaves room for further variations among climate models. Altogether, these alternative choices to be made are the reason why there is a large variety of quite different rather than a single climate model.

### *Retrodictions of current climate*

Before assessing climate forecasts, we should notice that this task is hampered by problems very similar to those we have encountered when evaluating the performance of macroeconomic forecasts. Which variables should be compared — temperature, precipitation or sea level change, global or regional values? What forecast horizon should be considered? Which error measures? ... As in the macroeconomic case, we risk to get lost in dimensionality. The IPCC (2001*b*) notes that "it has proved elusive to derive a fully comprehensive multidimensional 'figure of merit' for climate models" (p. 475). But in contrast to the macroeconomic case and worst of all, there is no survey of climate predictions yielding a forecast record which could be evaluated. That is the reason why the evaluation of climatology's predictive capacity has to be an evaluation of climate retrodiction.

Can climate models successfully retrodict the current global climate? A first way to answer this question consists in calculating and comparing the RMSE of

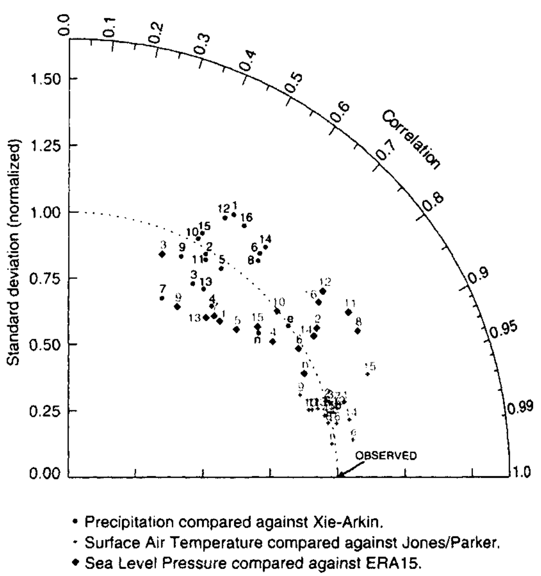



Figure 4.10: 16 climate models in a Taylor diagram; the linear distance between a model-point and the point of reference (OBSERVED) equals the corresponding RMSE. (Source: IPCC, 2001*b*, p. 482)

the climate models' retrodictions[13] for several variables. Figure 4.10 plots the forecast's normalized standard deviation and the forecast-observation correlation for (i) the global mean precipitation, (ii) surface air temperature and (iii) sea level pressure of 16 climate models in a polar coordinate system. The presentation in a so-called Taylor diagram ensures that the linear distance between the model-point and the observation-point $(1, 1)$ equals the forecast's RMSE.[14]

According to figure 4.10, the RMSEs are highest for precipitation forecasts and lowest for surface air temperature predictions. The IPCC (2001*b*) additionally notes that pooling of forecasts seems to improve their performance.

Another way to assess the predictive potential of climate models consists in asking whether they successfully retrodict typical evolutionary *patterns* of our recent climate. According to the TAR, a few climate models are able to correctly retrodict the global mean temperature rise during the 20th century. Figure 4.11 illustrates this observation by plotting the forecast of a model under different boundary conditions compared to observation (Parker/Jones). This kind of

[13] Conveniently referred to as forecasts, too.

[14] More precisely, the linear distance between a point representing a model and the point of reference equals the part of the RMSE that is not due to bias, divided by the predicted variable's standard deviation, see also a detailed proof in the appendix.

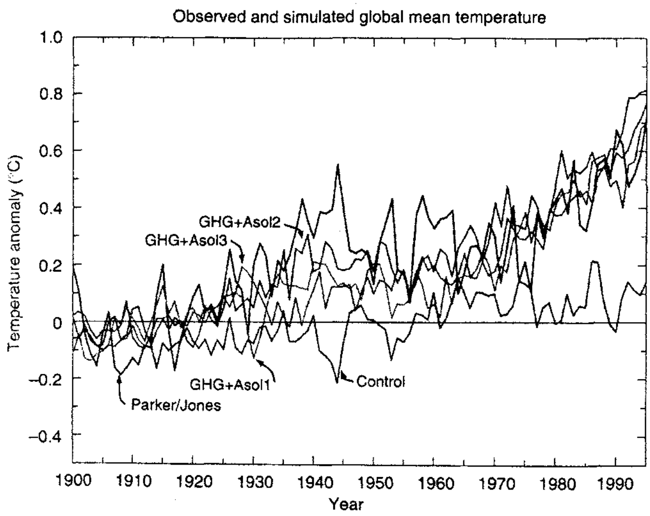



Figure 4.11: Observed (Parker/Jones) and modeled global annual mean temperature anomalies in the 20th century. The alternative boundary conditions assume GHG-emissions increase only (control) and GHG-emissions increase plus an increase in different aerosol concentration (GHG+Asol1-3). (Source: IPCC, 2001*b*, p. 497)

qualitative evaluation can also be applied on smaller time scales with regard to natural climate phenomena such as monsoons, and climate oscillations. The TAR is fairly vague on this issue stating that (i) some models can predict certain climate phenomena, (ii) there has been improvement since the Second Assessment Report (SAR) from 1995 and (iii) a systematic evaluation remains yet to be accomplished.

We have to ask critically whether the above quantitative and qualitative evidence warrants the conclusion that there is some kind of predictive success in climatology. First of all, the TAR does not present the reported facts as an evidence for successful retrodiction but rather for the successful *reproduction* of climate-data. This leads us to the question whether the retrodictions are really out-of-sample forecasts. If not, i.e. if the retrodictions are (indirectly) derived from what they were supposed to forecast, namely the current climate, they do not warrant the model's predictive reliability. Yet this is exactly the case as becomes evident when reading the TAR carefully,

> Recent atmospheric models show improved performance in simulating many of the important phenomena, compared with those at the time of the SAR, by using better physical parametrizations [...] (p. 506)

Hoewever, a certain parametrization is better than another only insofar as it results in empirically more adequate simulations. Therefore, data of today's climate is

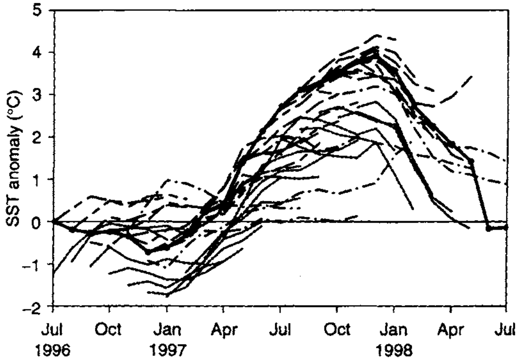



Figure 4.12: Surface sea temperature anomalies due to El-Niño, predictions (dash) and observation (solid). (Source: IPCC, 2001*b*, p. 504)

used to improve the parametrizations of and thus to construct climate models. The in-sample character of the retrodictions of 20th century climate becomes even more apparent when considering so-called flux-adjustments. Until recently all coupled climate models suffered fatal "climate drift": Not only did they diverge from the actual climate when being initialized with current conditions but they also yielded physically impossible simulation results. The only way to get rid of this problem was to introduce flux adjustments that modify energy and mass exchanges between the components. Yet these mathematical adjustments did not correspond to any real physical processes. In addition, the ad hoc adjustments were of the same size as the adjusted fluxes themselves! During the last few years, however, the evolution of climate models, notably new parametrizations and higher resolutions, brought about climate models without flux-adjustments whose simulation results were at least not absurd. However, the problem of climate drift is not resolved fully. All this shows that GCMs are constructed and continuously modified so as to fit the 20th century climate-data. A replication of this data may thus not count as an evidence for predictive performance.

What has just been concluded is qualified by a single exception that can be found in the TAR: the record of predictions of the El-Niño-Southern-Oscillation (ENSO). El-Niño is a regional climate phenomenon that periodically causes increased surface sea temperature in the tropical pacific. Several Institutes tried to predict the 1998 El-Niño. Figure 4.12 compares the predicted with the observed temperature anomalies. The forecasts had been revised several times during the whole period so that we do not know the exact forecast horizon — all we can say is that it was relatively short. Still, considering figure 4.12, we may doubt the TAR's conclusion that "current models can predict major El-Niño events with some accuracy". There is an illuminating analogy that can be drawn to economic forecasting: Purely statistical models (using time-series methods) can predict El-Niño with more accuracy than GCMs. Yet, the latter are performing better than simple climate models. So, even though the El-Niño forecasts represent real out-of-sample predictions, they do not provide conclusive positive evidence for climate models' forecast performance.

*Retrodictions of past climates*

Climatologist are quite aware that being able to successfully fit models to the data is no good reason to believe that the models are predictively successful: "Accurate simulation of current climate does not guarantee the ability of a model to simulate climate change correctly" [TAR, p. 493]. That is why it becomes necessary to assess retrodictions of past climates. Climate scientists have chosen two different climatic situations for this evaluation: (i) the mid-Holocene about 6,000 years before present (because of the different angle of the axis of earth-rotation) and (ii) the last glacial maximum around 20,000 years ago.

Regarding the mid-Holocene, the climate models have made correct directional forecasts of the following major climatic differences compared with current climate, but have largely underestimated them quantitatively: the northward displacement of the desert-steppe transition, the increased number of lakes in the Sahara, the northward shift of the Arctic tree-line and, finally, a northern expansion of the African monsoon. As to the last glacial maximum, the TAR is less informative: At least some models seem to produce realistic results. But inconsistencies with palaeo-data are used to infer — not that something is wrong with the model, but — that the estimations of historical sea surface temperature have to be revised. This perfectly illustrates the TAR's earlier statement: "We recognize that, unlike the classic concept of Popper, our evaluation process is not as clear-cut as a simple search for 'falsification'." (p. 474)

Does all this represent evidence for the predictive performance of climate models? Again, doubts arise. First of all, the correspondence of the model-data with the observations is even less obvious and much less documented than for the current climate. Secondly, it is not clear whether climatologist use the past climate solely as a reference for out-of-sample retrodictions or whether the assessment of these retrodictions are once again used to modify the models in certain aspects. For example: "The [...] simulation shows that combined feedbacks between land and ocean lead to a closer agreement with palaeo-data" [TAR, p. 495]. This information will certainly be used to improve future climate models — indeed, why should climatologists not exploit this piece of information? Still, by doing so,

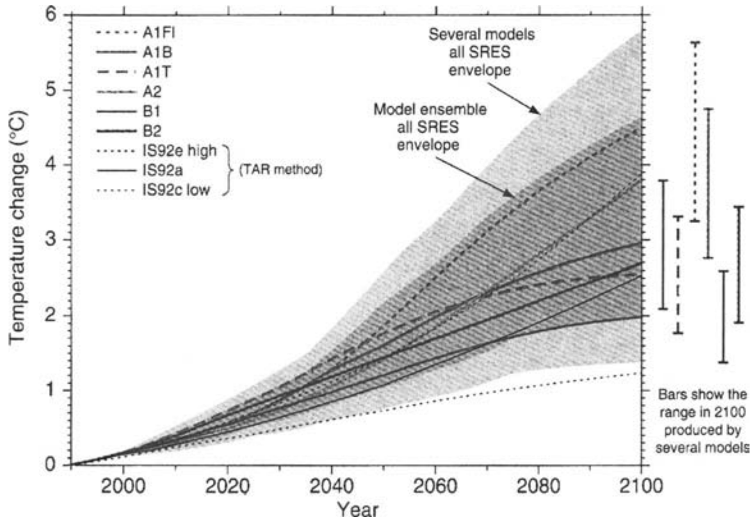



Figure 4.13: Global mean temperature change in the 21st century as predicted by several models and for different scenarios. (Source: TAR, p. 555)

they invalidate the evidence such retrodictions could have represented for climatology's predictive success.

*Projecting future climate*

Given our previous investigations, neither the simulation of current climate nor the retrodiction of past climates should count as a predictive success of climatology. Without a sample of past predictions and merely lots of predictions about the next century, there seems no other way to assess the climate models' forecast performance. Yet that is not correct, for we might outfox the future. Since forecast performance depends on a forecasts' accuracy as well as its precision and as a forecast sequence is the more precise, the less the individual forecasts diverge, we can anticipate forecast performance by examining a forecast record's variability. The more the forecasts disagree, the higher the (future) forecast error.

By the very nature of GCMs, climate predictions rely heavily on the co-prediction of the model's diverse boundary conditions. The TAR refers to a complete set of such boundary conditions as a scenario. Now, a whole system of different scenarios which has been developed under the auspices of the IPCC classifies different possible scenarios according to four main storylines (A1, A2, B1, B2) that differ relating to future global growth, speed of economic convergence and world population. A storyline may contain scenario groups that are characterized by cer-

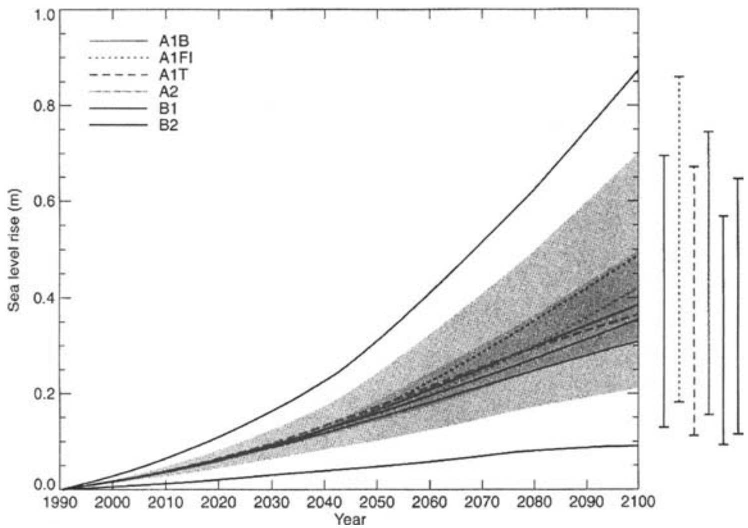



Figure 4.14: Global sea level rise in the 21st century as predicted by several models and for different scenarios. (Source: TAR, p. 671)

tain assumptions on technological development and policy measures. Altogether, there is a total number of 35 different scenarios.

Given the need for simplification in climate models and the diversity of the underlying scenarios, one could expect a fairly high degree of disagreement between different climate predictions. Figures 4.13 and 4.14 show the temperature and sea level change during the 21st century as predicted by a lean climate model that is tuned to different GCMs. Each dashed curve represents the categorical forecast based on one specific scenario and averaged over all the different models. The forecasted temperature rise by 2100 for example varies between 2°C and 4.5°C for the different scenarios. Yet forecasts not only disagree because of the alternative scenarios, but also because of the variety of models. Each vertical bar right to the diagram illustrates the disagreement between the forecasts based on one and the same scenario. In the case of temperature change, this model-induced disagreement is as high as the scenario-induced uncertainty. For sea level change, however, the disagreement due to the different models even largely exceeds the disagreement induced by the alternative scenarios. Altogether, the disagreement among forecasts is enormous; forecasts of temperature rise vary for example between 1.5°C and 5.8°C. Because of this imprecision, climate forecast performance has to be considered as poor.

What is the lesson to be learned from all this? The TAR concedes that the future climate is not predictable: "Because scenarios by their very nature should not be used and regarded as predictions, the term 'climate projection' is used in this Report." (p. 95) What this statement refutes is only the idea of categorical predictability. The extreme disagreement, however, between models suggests that the climate is not even conditionally[15] predictable.

The last chapter of the TAR contains a critical discussion of the predictability question and it plainly asserts "that the long-term prediction of future climate states is not possible. The most we can expect to achieve is the prediction of the probability distribution of the systems future possible states by the generation of ensembles of model solutions." (p. 774) Whether this is really possible will be discussed later in chapter 9. At this point, we simply note that the TAR does not provide any positive evidence for successful density forecasting in climatology, either.

## Main observations

**Observation 12 (Progress in depth)** *Experimental thermodynamics is characterized by predictive progress in depth that stems from improved experimental techniques.*

**Observation 13 (Progress in scope)** *Experimental thermodynamics is characterized by predictive progress in scope that stems from improved experimental techniques.*

**Observation 14 (Limits of forecasting)** *The limits of experimental techniques, and not the limits of theoretical understanding, are the limits of forecasting in experimental thermodynamics.*

**Observation 15 (Failure of climatology)** *There is no evidence for any successful forecast performance of climatology, neither with regard to absolute, nor conditional nor probability forecasts. Highly imprecise forecasts point to a poor predictive performance.*

**Observation 16 (Problems of forecast evaluation)** *Most of the problems an evaluation of so-called climate predictions faces are similar to those of macroeconomic forecast evaluation.*

---

[15] i.e. given a certain scenario.

# Part II

# Understanding the predictive limits of economics

# Chapter 5

# Predictability of complex systems


## Summary

The fact that the economy is a complex system explains some of the observations made in the previous part. Besides (i) yielding an argument from external effects which shows why macroeconomic forecasts generally fail, it gives rise to explanations of (ii) the more particular fact that directional forecasts do not perform any better than point forecasts, and (iii) the marginal role of macroeconomic theory for economic forecasting.


## 5.1 Explicating complexity

Complexity is not a property of a system as such. For what is complex is a system *under a certain description*. In addition, it is a comparative concept: A system may be more, less or equally complex than another. Complexity depends on the number of elements of the system, the number of its properties and the number of relationships between these elements or properties. We say that a system is simpler than another if it is less complex. These explications leave room for indecisive cases where it is not determined whether a system is more or less complex than another, for instance when one system is composed of more different elements but characterized by less different properties than the other. This, however, does not pose any difficulties for the following reasonings.

Conceiving a system mathematically yields more explicit indicators for complexity. A system is typically described mathematically by stating the functional relationships, $f_i$, which hold between the system variables, $x_i$, and involve certain parameters, $\alpha_i$.[1] Such a formal model has the shape:

$$\begin{aligned} f_1(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_m) &= 0 \\ f_2(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_m) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_m) &= 0 \end{aligned} \tag{5.1}$$

This is in fact the most general way to state a quantitative model at all. The degree of complexity can now be considered as a function of the numbers $n$ and $m$ as well as the average number of parameters altering the functional relationships. This definition applies no matter whether the variables $x_i$ are state variables, $l$-dimensional vectors, time derivatives of certain variables, or statistical properties as mean or variance.

Let us have a look at some examples. If the above definitions are applied to the systems we have studied in part 1, our intuitive judgments concerning their complexity are largely confirmed. Coupled atmosphere-ocean general circulation models (AOGCMs) are for instance more complex than AGCMs which in turn are more complex than simple energy balance models consisting only of a handful of equations. The econometric models used in forecasting institutes as for example the DRI/McGraw-Hill model are more complex than the simple model introduced in section 2.3. Finally, to give a non-formal example, the system that Adam Smith describes in *The Wealth of Nations* on hundreds of pages is more complex than the non-formal statement of the Ramsey-Cass-Koopmans model by Romer (1996, p. 39),

> There is a large number of [identical] firms. [...] The firms hire workers and rent capital in competitive factor markets, and sell their output in a competitive output market. The firms maximize profits. They are owned by households, so any output they earn accrue to the households. [...]
>
> There is also a large number of identical households. The size of each household grows at [constant] rate [...]. Each member of the household supplies one unit of labor at every point in time. In addition, the household rents whatever capital it owns to firms. [...] The household divides its income at each point in time between consumption and saving so as to maximize lifetime utility.

[1] Although this is by no means the only type of formal mathematical description of a system, it is by far the most (maybe even the only) relevant in economics and physics.

## 5.2 External effects

The above explications enable us to reconstruct a familiar argument against the possibility of macroeconomic forecasting called the argument of external effects. The state of the system (5.1) at time $t$ given by the values of $\mathbf{x} := (x_i), i = 1 \ldots n$ clearly depends on the values of the parameters $\mathbf{a} := (\alpha_j), j = 1 \ldots m$ at time $t$. If someone wants to derive a prediction about the future state of the system, ${}_{t-1}\mathbf{x}_t^F$, with (5.1) as predictive model, the future values of the parameters ${}_{t-1}\mathbf{a}_t^F$ must therefore be co-predicted. But even if (5.1) does not serve as a predictive model, a forecast ${}_{t-1}\mathbf{x}_t^F$ entails at least negative co-predictions of the parameters. For although a certain prediction ${}_{t-1}\mathbf{x}_t^F$ might be compatible with several future parameter sets, there is a large number of parameter combinations that exclude, together with (5.1), the forecast. So, by whatever procedure the forecast is established, it entails that certain parameter values will not be realized in the future. In other words, any prediction of a system's state entails a co-prediction of its future boundary conditions.

Now, the co-predicted states are the more numerous, the more complex the predicted system. Forecasts of very complex systems commit the forecaster to predict a large number of further (boundary) states. And the correctness of the original prediction depends on whether the co-predictions turn out to be correct or not. This clearly gives rise to a major problem when predicting complex systems. And it explains the observations we have made in part 1. With respect to economic forecasting, the argument was elaborated by Morgenstern (1928). He conceives the economy as a very complex system that intensively interferes with political, social or natural phenomena. Having noted that the main idea of his reasoning is quite popular and maybe even trivial, Morgenstern says,

> The sentence: Economic forecasting that makes use of the means provided by economic theory is principally impossible has to be interpreted [...] as follows: Even if there were a positive theory of economic forecasting, what we deny, economic prediction as such would still be insufficient since economic processes are co-determined by other factors, other types of behavior that would have to be co-predicted, but cannot be predicted. In other words: A prediction of economic processes is *eo ipso* a co-prediction of other types of behavior that are somehow reflected quantitatively in economic behavior. Yet that goes in any case beyond the realm of possibility of economic theory, no matter how brilliant the theory of economic forecasting. (p. 115)

The argument of external effects can now be stated concisely,

(1) The economy is a complex system that largely interacts with political, social or natural phenomena.
(2) Any prediction of a complex system entails (positive or negative) co-predictions relating to the phenomena the system interferes with.
(3) *Thus:* Macroeconomic forecasts entail (positive or negative) co-predictions of political, social and natural phenomena.
(4) Forecasts of political, social and natural phenomena fail.
(5) *Thus:* Macroeconomic forecasts fail.

However, what is true for the economy as a complex system, is also valid for complex natural systems, as for instance the earth climate. We saw that climate predictions, e.g. of global mean temperature rise, heavily rely on co-predictions of related parameter and boundary conditions such as the underlying scenarios and the specific parametrization of the GCM. In contrast to some economists and as already mentioned earlier (see p. 78 above), climatologists concede that the co-prediction of these states is not possible and consequently propose to modestly label their predictions "projections".

## 5.3 Directional and non-directional forecasts

In part 1, we have observed that macroeconomic trend forecasts are as likely to fail as quantitative forecasts (Observation 3). This seems to be counter-intuitive. And many economists, whilst admitting that macroeconomic point forecasting is hardly feasible, insist on the possibility of directional forecasting. I suppose that something like the following argument is what feeds these intuitions,

(1) If forecasts of some type F are deduced from weaker assumptions than forecasts of another type G, then F-forecasts are less likely to fail than G-forecasts.
(2) Point forecasts are based on assumptions on the precise values of certain variables (quantitative predicting data) whereas directional forecasts merely require assumptions on the variables' signs (qualitative predicting data).
(3) Assumptions on the signs of some variables are weaker than those on their precise values.
(4) *Thus:* Directional forecasts are less likely to fail than non-directional ones.

Given our observation, something must be wrong with this argument. Seeing what is wrong will enable us to explain why directional forecasts do not perform any better than point forecasts.

In the following I will argue that premiss (2) does not hold in sufficiently complex systems. This result will be obtained by making use of the methodological framework Paul Samuelson has developed in his *Foundation of Economic Analysis* for exactly the opposite purpose (Samuelson, 1963). Samuelson tried to show that it is a reasonable as well as an achievable aim of economics to establish what he called tendency laws, and he investigated the general conditions which must be satisfied in order to derive such laws. A tendency law is a general relationship which holds in a specific system and relates the sign of change in a certain parameter $\alpha_l$ to the sign of change in some variable $x_k$, i.e. it specifies the sign of $dx_k/d\alpha_l$. Accordingly, conditional directional forecasts such as 'If $\alpha_l$ is increased, $x_k$ increases, too.' can directly be derived from it. As the antecedent conditions of these conditional directional forecasts are non-numerical, a tendency law which is itself not derived from numerical assumptions makes it possible to derive categorical directional forecasts (by co-predicting the antecedent conditions) which are based on qualitative predicting data only.

But in sufficiently complex systems, the sign of $dx_k/d\alpha_l$ cannot be derived from purely qualitative information only. To see why, reconsider the mathematical description (5.1) of a system; differentiating every equation in the general case with respect to $\alpha_l$ yields a linear equation system with $n$ variables ($dx_j/d\alpha_l$, $j = 1 \ldots n$) and $n$ equations,

$$\begin{array}{ccccc}
\frac{\partial f_1}{\partial x_1}\frac{dx_1}{d\alpha_l} & + \ldots + & \frac{\partial f_1}{\partial x_n}\frac{dx_n}{d\alpha_l} & = & -\frac{\partial f_1}{\partial \alpha_l} \\
\frac{\partial f_2}{\partial x_1}\frac{dx_1}{d\alpha_l} & + \ldots + & \frac{\partial f_2}{\partial x_n}\frac{dx_n}{d\alpha_l} & = & -\frac{\partial f_2}{\partial \alpha_l} \\
 & \vdots & & & \\
\frac{\partial f_n}{\partial x_1}\frac{dx_1}{d\alpha_l} & + \ldots + & \frac{\partial f_n}{\partial x_n}\frac{dx_n}{d\alpha_l} & = & -\frac{\partial f_n}{\partial \alpha_l}
\end{array}$$

The solution of this equation system provides the derivative, $dx_k/d\alpha_l$, we are looking for, and can be represented, according to *Cramer's Rule*, as the quotient of two determinants, $\det A^* / \det A$, where (i) $A$ is the matrix of all partial derivatives $[\partial f_i/\partial x_j]_{ij}$ and (ii) $A^*$ is obtained by substituting the vector $(-\partial f_i/\partial \alpha_l)_i$ for the $k$th column of $A$. Now, let us assume maximum qualitative knowledge, i.e. the knowledge of all signs of the matrix-elements. Since we would have to use *Leibniz' Formula* with its $n!$ terms to calculate the sign of a determinant in the absence of quantitative knowledge, we would only be able to calculate the sign of the derivative, $dx_k/d\alpha_l$, we are interested in if these $n!$ terms had the same sign for each of the two determinants. As Samuelson himself points out, this is of utmost unlikeliness for large $n$. Here, I claim more than Samuelson: It is not only unlikely, it is impossible for $n \geq 3$!

To prove this, we will consider a $3 \times 3$ matrix $B_{ij} = [b_{ij}]$ with $i, j = 1, 2, 3$ in a first step. We assume that $b_{ij} \neq 0, \forall ij$. According to Leibniz' Formula, the determinant can be calculated by summing $3! = 6$ terms,

$$\begin{aligned} \det(B_{ij}) = & \quad b_{11}b_{22}b_{33} + b_{12}b_{23}b_{31} + b_{13}b_{21}b_{32} \\ - & \quad b_{11}b_{23}b_{32} - b_{12}b_{21}b_{33} - b_{13}b_{22}b_{31}. \end{aligned}$$

All of these six terms, each being a product of three coefficients, will only be of positive sign if (i) an even number of coefficients in each of the three terms in the upper row and (ii) an odd number of coefficients in each of the three terms in the lower row are of negative sign. This would imply that there were an even (total) number of negative coefficients in the upper row and an odd (total) number of negative coefficients in the lower one. But that is impossible since each coefficient occurs exactly once in the two rows. An analogous reasoning shows that not all of the six terms can be of negative sign. Thus, we are not able to determine the sign of a $3 \times 3$ matrix qualitatively. Now, let $A$ be a $n \times n$ matrix, again, none of the entries equal to zero. Making use of *Laplace's Theorem*, the terms of the Leibniz' Formula can be arranged in $n!/6$ groups: Each group is the Leibniz' sum of a $3 \times 3$ submatrix $A^k$ $(k = 1 \ldots n!/6)$ of $A$ multiplied by a certain factor and is therefore composed of six terms. Since we have just derived that these six terms cannot be of the same sign, it is shown by this that the terms of the Leibniz Formula are not of equal sign for any $n \geq 3$.

Hence, in sufficiently complex systems, qualitative knowledge (predicting data) is not enough to derive tendency laws and directional forecasts. Information about the future trends of all relevant variables is not sufficient to determine the future trend of the predicted variable. Quite the opposite, its future trend depends on the precise values of the predicting variables and parameters. This said, it becomes clear that premiss (2) of the above argument is false; the intuition that directional forecasting is easier than point forecasting is unfounded. Moreover, the failure of point forecasts to some extent even entails the failure of directional forecasts in complex systems: A directional forecast relies on the same predictive model and the same (quantitative) predicting data the point forecast is derived from. If the latter fails, the former relied (for modus tollens) on false assumptions and is thus likely to be incorrect, too.

## 5.4 Theory independence of macroeconomic forecasting

We have found in part 1 that macroeconomic forecasts are not primarily based on theoretical models. Economic theory is only used in so far as some huge econometric models are applied; yet these do not have much in common with the highly idealizing models of modern macroeconomic theory. Why is economic forecasting so disconnected, so independent from macroeconomic theory?

The highly idealized models of macroeconomic theory are characterized by relatively low complexity, i.e. they contain relatively few variables and parameters. Yet this is something they have in common with classical physical models whose simplicity Cartwright (1983*a*) illustrates with the laws covering the diffusion of liquids in a mixture,

> The situation is this. For several fluxes $J$ we have laws of the form,
>
> $$\begin{aligned} J_m &= g_1^*(\alpha_m) \\ &\vdots \\ J_q &= g_n^*(\alpha_q). \end{aligned}$$
>
> Each of these is appropriate only when its $\alpha$ is the only relevant variable. For cross-effects we require laws of the form
>
> $$\begin{aligned} J_m &= g_1(\alpha_m, \ldots, \alpha_q) \\ &\vdots \\ J_q &= g_n(\alpha_m, \ldots, \alpha_q). \end{aligned}$$

Cartwright notes that we have laws of this latter kind just in very few cases and that there is no universally applicable procedure to combine laws of the simple kind in order to obtain laws that describe a complex system adequately. While Cartwright's example of cross-effect laws turns out to be a special case of the equation system (5.1), her simple law example represents a special case of the system,

$$\begin{gathered} f_1^*(x_1, \ldots, x_n, \alpha_1) = 0 \\ f_2^*(x_1, \ldots, x_n, \alpha_2) = 0 \\ \vdots \\ f_n^*(x_1, \ldots, x_n, \alpha_n) = 0 \end{gathered} \tag{5.2}$$

System (5.1) is clearly more complex than (5.2) where each functional relationship is altered by *one* parameter only. Cartwright asserts that physical laws generally take the form (5.2).[2] Samuelson does not consider it as problematic that this applies to economics as well, since

[2] Von Hayek (1972) does not only consider this as a matter of fact but suggests to use it as a criterion to define physical science: "The non-physical phenomena are more complex because we refer to what can be described by relatively simple formulas as physical." (p. 13)

> the assumption of this hypothesis [that each functional relationship is altered but by one parameter] does not involve any serious loss of generality and still includes the vast majority (in fact, it is hard to find exceptions) of relationships contained in current economic theory. (Samuelson, 1963, p. 33)

If Samuelson simply wants to remind us that economic models are generally of type (5.2), he is probably right, but if he intends to claim that economic phenomena can generally be described adequately by such models, he is certainly wrong. Cartwright's observation suggests that classical physics and economic theory can hardly be applied to complex situations at all. This is nicely illustrated by our climate models. GCMs are clearly special cases of (5.1). And they are obviously not simply composed of natural laws nor are they strictly reducible to thermodynamics or chemistry. On the contrary, some of them are even incompatible with basic physical laws. The ad-hoc adjustments of energy and momentum fluxes that had to be implemented in some AOGCMs in order to ensure the model's stability implied that, a priori, the model did not comply with fundamental conservation principles, i.e. the conservation of mass and energy!

In order to describe a complex system adequately by economic or physical theory, it has to be manipulated so that it can be subsumed under a model of type (5.2). Such a manipulation typically consists in shielding (parts of) the system against any kind of external influence, reduces the system's complexity, and eventually transforms it into a simpler one. Or, in the terminology of Cartwright (1999), complex systems may be rearranged to nomological machines, i.e. an appropriate arrangement of certain factors in a sufficiently shielded environment that gives rise to regularities of the type (5.2).

This said, it seems as if complex systems on the one hand and theoretical laws on the other just do not match. The models that attempt to describe the former adequately cannot be reduced to the latter. And systems which are governed by the latter are, opposed to the former, simple, well-shielded and controlled. However, it is not merely a contingent matter of fact, possibly a cognitive shortcoming, that scientific laws are simple. There is a methodological reason for the fact that our theoretical laws don't apply to complex systems, as we shall see in the following.

The reconstruction of Samuelson's account has made plain that simplicity is necessary in order to derive tendency laws in a purely qualitative way. If we acknowledge that qualitative understanding means to derive such general, stable qualitative relationships, it follows that theorists will consider simple systems only as long as they aim at a qualitative understanding of the studied system.[3] In ad-

[3] I do not mean to defend this goal; I merely say that because scientists have adopted this goal, they have developed theories that do not describe complex systems adequately.

dition, Samuelson has shown that simplicity, together with a second condition, is also *sufficient* to deduce tendency laws. These conditions are:

1. The system prevails in an equilibrium so that the function

$$F(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_m)$$

   with $\partial F/\partial x_i = f_i$ is maximal. The equations that state the corresponding first order condition form an equation system.

2. Each functional relationship of the first order derivative system is altered by one parameter only. Hence, the equation system is of the kind (5.2)!

However, Samuelson's two conditions are not equally important for deriving tendency laws. While the first condition might be dropped, condition (2), the simplicity-condition, is crucial. Not assuming that the system prevails in equilibrium gives rise to the necessity to study its dynamical behavior. As a result, the variables typically become first order time derivatives of state variables. Let us examine the above mentioned Ramsey-Cass-Koopmans (RCK) model as a final case study[4] in order to illustrate the importance of the simplicity condition even in dynamical analysis.

If the assumptions of the RCK model are transformed from ordinary language (see p. 82 above) into formal functional relationships, the following two equations can be derived,

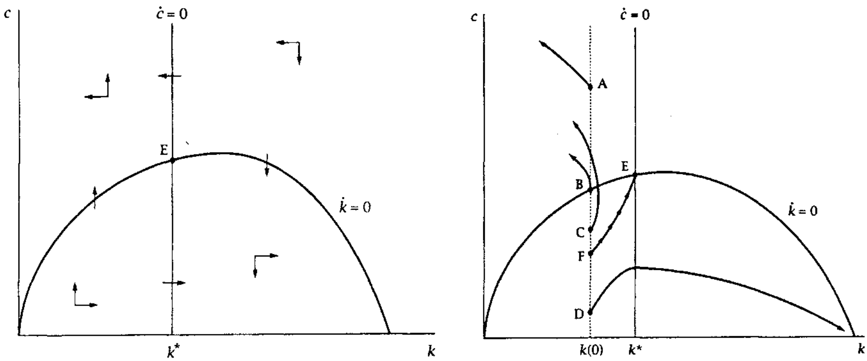
$$\begin{aligned} \dot{c}(t) &= \frac{c(t)}{\theta}(f'(k(t)) - \rho - \theta g \\ \dot{k}(t) &= f(k(t)) - c(t) - (n+g)k(t) \end{aligned}$$

where $c(t)$ is consumption, $k(t)$ capital stock, $f(*)$ is the production function of capital and $f'(*)$ the marginal productivity of capital, all functions of time $t$ and per unit of effective labor. The parameters are productivity growth $g$, population growth $n$, discount rate $\rho$ (the greater $\rho$, the less the households value future consumption relative to current consumption) and the coefficient of relative risk aversion $\theta$. These equations can also be written as,

$$\begin{aligned} f_1(\dot{c}(t), \dot{k}(t), c(t), k(t), \rho, \theta, g) &= 0 \\ f_2(\dot{c}(t), \dot{k}(t), c(t), k(t), n, g) &= 0 \end{aligned} \quad . \tag{5.3}$$

for some $f_1, f_2$.

[4] Following the textbook account in Romer (1996).



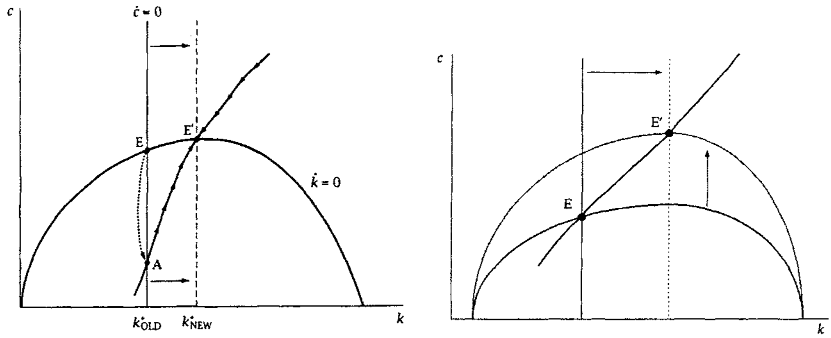

(a) Setting up the phase diagram. (b) Ruling out contradicting time paths.

Figure 5.1: The phase diagram of the RCK model. (Source: Romer, 1996, p. 48-49)

Qualitative statements about the economy's behavior are derived by the following procedure. The first step consists in drawing the phase diagram. This not only demands specific qualitative knowledge of the functional relationships but for obvious reasons also requires that the system is not composed of more than two differential equations. This is a first constraint regarding the system's complexity. Figure 5.1a plots the phase diagram of the Ramsey-Cass-Koopmans model.

Still, the phase diagram depicts infinitely many possible time paths the system could evolve upon depending on its initial position and the numerical values of the parameters. In a second step, thus, as many time paths as possible must be ruled out on the basis of the model's assumptions. This is illustrated in figure 5.1b. For any initial capital stock $k(0)$, there is but one corresponding consumption level that does not bring about time paths that contradict the model's assumptions. It can subsequently be shown that the system necessarily prevails on the saddle path $FE$ and converges against the equilibrium $E$. The reasoning thus nicely exemplifies what Samuelson has called the *Correspondence Principle* between comparative statics and dynamics: Instead of assuming that the state variables prevail in a stable equilibrium position, the dynamical analysis tries to derive tendency laws by assuming that the model-economy is in 'stable' motion.

Given these results, we are prepared to investigate, in a third step, the effects of shifts in certain parameters. A drop in the discount rate $\rho$, for instance, alters but the first equation of the system (5.3) and causes the vertical $\dot{c} = 0$ line to shift to the right in the phase diagram 5.4a. Since we have derived that the model is always on the saddle path, it follows — assuming that only consumption and not the capital stock can be adjusted instantaneously — that consumption drops sharply to $A$ on the new saddle path and subsequently adjusts to the new equilibrium $E'$.




(a) Definite effect. (b) Ambiguous effect.

Figure 5.2: Effects of parameter shifts in the RCK model. (Source: Romer, 1996, p. 54)

And this is exactly what we have tried to derive: a conditional directional forecast.

However, we were lucky that $\rho$ alters but one single equation of (5.3). If this had not been the case, the results of a change in the discount rate could have been ambiguous. Consider for instance a decline in productivity growth $g$, a parameter that affects both equations in (5.3). As for the drop in $\rho$, the $\dot{c} = 0$ line is displaced to the right. Simultaneously, the $\dot{k} = 0$ curve is vertically expanded. Thus, it depends on the quantitative extent of these transformations whether consumption immediately falls or rises. Figure 5.4b shows the case where the effects neutralize each other and consumption neither drops nor rises immediately. In addition, it is not determined from a purely qualitative point of view whether the model adjusts to a higher or a lower consumption rate in the long run, either. Thus, this case study confirms: As soon as the second of Samuelson's two conditions is not satisfied, qualitative forecasts can no longer be derived qualitatively. Hence, insofar as *qualitative* understanding is what is aimed at, science *has to* deal with simple systems of a very low degree of complexity.

To round up the whole argumentation, we must recall that the economy as we use to conceive it in ordinary language — talking about job opportunities, firms, new products, stock markets, shopping centers, taxes, ... — is anything but a simple, sufficiently shielded system of low complexity. So, given the arguments of this chapter, it is fairly comprehensible from this every day point of view that (i) the whole enterprise of forecasting the economy fails because of the problem of external effects, that (ii) directional forecasts are as unlikely to be correct as point forecasts and, finally, that (iii) economic theory is — given the aim of qualitative understanding — necessarily simple and therefore of little use to macroeconomic forecasting. This holds in analogy for the earth climate and other natural systems we ordinarily conceive as very complex.

# Chapter 6

# The quality of data


## Summary

Data cannot possibly be compared with the 'true data' in order to assess its quality. Instead, its quality is determined by its precision as well as by the evidence in favor of its underlying assumptions. In general, forecasts are negatively affected by inferior data-quality. Both macroeconomic and climate-data suffer from typical problems of aggregation and their quality, in terms of precision as well as of evidence for underlying assumptions, is poor. Thus, a considerable part of macroeconomic and climate forecast failure can be explained.


## 6.1 Explicating data-quality

Before explicating the notion of data-quality, let us fix the concept of data itself: Quantitative data consists in numerical values of a certain variable (GDP-data, CPI-data, ...); and "measurement" refers to the systematic generation of quantitative data. This said, it is worth noticing that the idea of data-quality is not only of particular importance in the context of forecast-evaluation. It also plays a major role in science in general, e.g. when certain results are refused by criticizing their data-basis. But what exactly do we mean by quality of data? What makes data *good* data? We faced a similar question right at the beginning of our enterprise when asking what forecast-quality consists in. In that case, however, we were in a far better position, since given the actual outcome of the predicted event we were able to calculate forecast errors and to relate forecast quality to these errors. In contrast, there is nothing like the actual or the true data with which we could compare the data under examination. For to say which data is true or 'more true' in order to specify the objective benchmark would presuppose an opinion about which data is of better quality — what obviously begs the question. Thus, we cannot evaluate data-quality the same way as we have assessed forecasts. Alternatively, we may judge data on the basis of its *precision* and of the *assumptions* which underlie its generation.

The first way to judge data, i.e. by its precision, consists in comparing it with other data that measures the same variable. In the forecast-case, precision was defined as volatility of forecast errors. In analogy, we may define precision of data as the volatility of a data-set quantifying the same variable and among others including the data under examination. Consequently, any data of the considered data-set has the same precision. To cite an example, the data which measured the latent heat of fusion became more and more precise during the last centuries, as described in chapter 4. The values obtained by Wilcke and Black in the second half of the 18th century varied by more than 5 cal/g. In the mid 19th century, data obtained via different measurements converged and differed only by around 1 cal/g. Finally, modern measurements only differ by less than a thousandth cal/g. These explications correspond with our ordinary intuition that the number of decimals indicate the precision of some data, assumed that only those decimals are given which actually agree for different measurements. For then, the less volatile (and thus the more precise) the data, the more decimals agree.

Although we typically compare data generated at the very same time in order to assess its precision, e.g. we repeat the measurement process several times and analyze the resulting data-set, it is equally possible, from the point of view just developed, to consider data-sets which are composed of data generated at very different times. Still, we tend to treat data that was generated at different times asymmetrically: The more recent data is typically considered to be of better quality and naturally serves as a benchmark. Such a practice may be justified in the light of corrections of the former measurement or other progress concerning the measurement-process. But it suggests that the notion of precision does not entirely grasp the idea of data-quality and that there is a further component of data-quality which involves the evaluation of data generation. This is the above-mentioned second way to assess data-quality.

More fundamentally, precision alone is not a satisfying criterion for data-quality because as a relative measure, characterizing some data compared to other data, it cannot take the possibility into account that all the data might be systematically wrong (biased). If a data-set is biased, for example because a certain heat-loss in the experimental setting is systematically underestimated, all the data may largely agree and hence be very precise despite of being — as we would still say — of bad quality. Since it is impossible to calculate the data's accuracy as we have done for forecasts, the only way to identify systematical errors is to evaluate the process of data generation and the assumptions this process is based on. The more the underlying assumptions and theories are confirmed, the better the quality of the generated data. A familiar problem occurs if the only evidence for the measurement-theory comprehends the measured data itself: The confirmation of the measurement-theory would then be invalid since circular. In this case, theory-ladenness has become fatal.

## 6.2 How does data-quality influence forecast-quality?

With these explications of data-quality at hand we now have to clarify in how far bad data-quality impairs forecast-quality and hence can explain forecast failure. Data-quality can affect forecast performance through two main mechanisms for data is required to derive (predicting data) as well as to evaluate (predicted data) a forecast.

*Forecast generation and data-quality*

Imprecise predicting data entails imprecise forecasts. This becomes clear when we do not only consider forecast errors (in line with the statistical stance developed in the introduction) but also the predicting data as a random variable. The larger the latter's variance, the larger the former's. The degree to which forecast precision is affected by data imprecision of course depends on the specific predicting model. Thus, if the predicting model is for instance a linear equation system, the forecast's precision linearly depends on the predicting data's precision.

Things are slightly more complicated for the second criterion of data-quality, namely the evidence for the underlying assumptions. For it is not generally true that a forecast is to fail the more likely the less justified the statements it is derived from.[1] The explanation of forecast failure by poor evidence for the predicting data's assumptions is hence not straightforward. However, we have introduced the assumptions' justification as a criterion for data-quality above because they seemed to be a potential indicator for the data's bias. The better the underlying assumptions are justified, the less likely a severe bias of the data. Biased predicting data, in turn, entails inaccurate forecasts since the data's bias propagates through the predicting model into the forecast. Consequently, the lack of evidence for the predicting data's underlying assumptions makes forecast inaccuracy more likely and can thus explain its failure.

*Forecast evaluation and data-quality*

Forecast evaluation makes use of data, namely by calculating the forecast errors. We shall now briefly discuss whether the forecast performance is affected by the

[1] As we will see in chapter 10.

| PREDICTING DATA | PREDICTED DATA | |
|---|---|---|
| | *bad quality* | *good quality* |
| *good quality* | GDP indicator forecast* | nomological machines |
| *bad quality* | econometric GDP forecast* | climate-model based sea-level rise projection* |

Figure 6.1: The data matrix, based on the results of this chapter. The failure of forecasts marked with a "*" can be explained by poor data-quality.

predicted data's quality, i.e. the data measuring the real outcome, too.

The imprecision of the predicted data has ambivalent effects on forecast performance. It might result in higher as well as in lower forecast errors, depending on which of the alternative data[2] is chosen in order to evaluate the forecast and to calculate its errors (e.g. whether to evaluate a forecast against the OECD- or the ILO-data, for instance). Such imprecision thus bears the risk that forecasters select that kind of data which minimizes calculated forecast errors and makes forecasts thus appear in a better light.

Weak underlying assumptions of the predicted data, on the contrary, seem to worsen forecast performance. To see why, consider a forecast, ${}_{t-1}X_t^F$, evaluated at time $t$. Now, all information available at $t-1$ is available at $t$, too. Or, the other way around, whatever information was not available or weakly confirmed at $t$, was not available or (possibly even less) weakly confirmed at $t-1$. Thus, if there is weak evidence for the assumptions underlying the measurement of the predicted variable at $t$, how much weaker are the assumptions underlying the predicting data at $t-1$! But weak evidence for the assumptions underlying the predicting data explains forecast failure as we have just argued. This said, poor quality of the predicted data rather than being a proper explanation merely indicates an explanation of poor forecast performance, namely the weak evidence for the predicting data.

### *Different types of data and the data-matrix*

Hitherto, we have discussed two types of data whose quality might affect forecast performance: the predicting and the predicted data. Anticipating the results of the following sections, this distinction enables us to arrange different kinds of forecasts in a matrix according to their corresponding predicted and predicting data's quality (figure 6.1). Poor performance of forecasts involving either predicting or predicted data of bad quality is thus, in line with the previous discussion, at least partially explained.

2 Recall that data imprecision means disagreement of alternative data.

| Scale | Entities | Epistemological access |
|---|---|---|
| macro | societies, atmospheres, galaxies | indirectly through aggregation |
| | | ↑ |
| meso | human beings, trees, spacehelmets | through direct observation |
| | | ↓ |
| micro | DNA, ozone molecules, neutrinos | indirectly through instruments |

Figure 6.2: The scale diagram, depicting our standard ontological picture.

## 6.3 General data-related problems in macro systems

The world-we-live-in is the world of tables, knives, sunflowers, trees, stones and things like that. Things we can see, touch, smell, taste or hear. The world-we-live-in consists of smaller and bigger things. However, most of us agree that there are very small things that are too tiny to be seen, touched, smelled, tasted or heard. They do not belong to to the world-we-live-in. Such things are for example the ozone-molecule or the HI-virus. And we need instrumental observation to infer their existence. Similarly, there are things that are too big to be seen, touched, smelled, tasted or heard. They do not belong to the world-we-live-in, either. The economy is an example for such a thing, the atmosphere of the earth another. These things are so huge that we can only grasp very small parts of them simultaneously by our senses. Their existence and their properties are inferred by the aggregation of direct observations in the world-we-live-in. Figure 6.2 summarizes this ontological picture, distinguishing the meso scale, i.e. the world-we-live-in, the macro scale and the micro scale. It is an over-simplified picture which might be contested in several ways, for instance because some properties of meso entities are inferred by successive instrumental observation and aggregation. However, it will serve its purpose here, namely pinpointing the problems related to data-quality in macro systems.

Since the economy is an example of a macro entity and properties of these entities are inferred by aggregation, we will now discuss typical problems of data-quality that arise when the measurement consists in aggregation. These problems relate to the assumption-aspect of data-quality and can be divided into two sorts: first, those problems that stem from incomplete meso information, and second, problems relating to the compatibility of the meso data.

Macro systems are described by a large number of observational (meso) statements. Measuring macro variables by aggregating such statements allows to describe the macro system in a few sentences. Such aggregation typically simply consists in summarizing meso data or in calculating its mean or other statistical properties. Now, problems might arise because of gaps in the data basis. In fact, the larger the system, the more likely it is that we do not possess a complete description of the system in meso terms. The sheer amount of information might not be manageable even with modern electronic data-processing. In this case, global variables must be calculated from an incomplete meso picture of the system. And this necessarily involves assumptions that specify how the gaps shall be filled. The assumption of homogeneity for example might serve this task: The missing data is assumed to be similar (in a certain respect) to the known meso data. In this case, variables are simply extra- (loose ends) or interpolated (gaps). Whether the assumption of homogeneity is justified or not must be judged on a case by case basis. The more diverse and heterogeneous the meso data, the less evidence there is in favor of the homogeneity-assumption and the more critically the aggregated data has to be judged.

Besides extra- or interpolation, gaps in the meso statistics might be filled by inference from other available meso information. This procedure consists in calculating the missing meso data from meso data of another type and is therefore based on assumptions relating the two types of data to each other. The quality of the aggregated data obviously depends on the evidence for these assumptions. Ideally, they are part of a well-confirmed scientific theory. Yet, they sometimes are really just ad hoc assumptions based on educated guess rather than on empirical research.

Meso data that serves as input of an aggregation not only should be as complete as possible but it must also be compatible. Consider a macro system that is composed of a large amount of distinct meso entities and let there be a property that can be quantified for each of these entities. Aggregation of the individual values of this property yields a global characterization of the macro system. But on which conditions is such a statement meaningful and informative? That depends on what we want to characterize, on the purpose the aggregated data should serve, the question the statement is supposed to answer. Let us consider a simple example: a rich garden with plenty of different trees and bushes, fruit of all kinds. Last year, each fruit-plant yielded a certain number of fruit (56 apples, 23 strawberries, ...). For each plant, the last year's yield of fruit is an appropriate indicator of the plant's harvest and allows for comparisons with former years. However, is it meaningful to summarize the individual yields to a global number? If it should serve as an indicator of the garden's overall harvest that allows for a comparison with the harvest two years ago, the answer is no, since higher numbers might simply be due to the currants and peas, understating the harvest of other less numerous but bigger fruit as melons or apples. Although all the meso data measures harvest, it is — relative to the purpose — not compatible with each other and should not be aggregated. To calculate a meaningful indicator of the garden's overall harvest, one either has to aggregate another meso variable or to manipulate the original data, for example by weighing it. If, however, the question is 'How often did I have to raise my arm to pick a fruit last year?', the simple aggregate might be quite appropriate.

This said, a necessary condition for aggregating meso data which quantifies a certain property P of meso entities is *comparability*, i.e. that the individual entities can be meaningfully compared with respect to the property P (e.g. this one is more P than that one) by comparing its quantitative values.[3] The general reason for this is that the basic calculation involved in aggregation is summing. This requires an addition to be defined on the meso data. So, the real values attributed to each meso entity cannot be thought of as belonging to different sets but must belong to one and the same set on which the addition is defined. This, however, induces an ordering on the set of entities with respect to the quantified meso property. Hence, if such an ordering is nonsense, so is aggregation, too. Whenever the meso data is not comparable but should nevertheless serve for an aggregation, it has to be manipulated. This for instance is the case when the meso data has been measured on different conditions that disturbed the obtained values. Such effects must be neutralized to ensure comparability.

We may conclude from the preceding paragraphs that the more diverse the meso data with respect to conditions and object of measurement, the more manipulation is needed. Any such manipulation involves certain assumptions that relate different meso variables to each other or that quantify the impact of different measurement conditions on the results. As in the case of an incomplete data basis, the quality of the macro data obtained by aggregation crucially depends on the evidence for or against these assumptions.

Such are the typical problems concerning data-quality in macro systems. Yet it cannot be established a priori that macro data is of bad quality. Instead, we have identified potential problems and will subsequently have to examine whether they actually arise in the case of macroeconomic and climate-data generation.

## 6.4 Quality of GDP-data

*Main principles of national accounting*

The gross domestic product (GDP) is one of the principal global variables that characterize an economy. It is supposed to measure an economy's total output. While national accounting serves several purposes, the calculation of GDP and

[3] Notice that one cannot sensibly compare the harvest of two different fruit plants by comparing the number of fruit yielded by each of them.

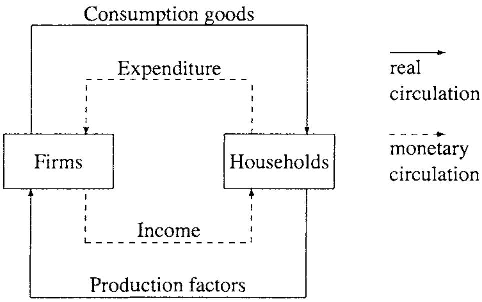



Figure 6.3: A simple model of economic circulation.

related aggregates such as the gross national income (GNI) is one of its major aims.

National accounting is based on the fundamental assumption that the economy can be described as a circulation.[4] Figure 6.3 shows the most simple economic circulation model that can be conceived and which consists of a real and a monetary circulation. The national accounts in general and the measurement of GDP in particular assume that each circulation is closed, i.e. that for any node, inflow equals outflow. It is assumed additionally that the value of the real flows is equal to the corresponding inverse monetary flows, e.g. the value of the consumption goods equals expenditure. Since, in our simplified model, the total output consists in the total of all consumption goods, there are four ways to measure GDP: by measuring the total value of produced goods and services (production side), by measuring the total expenditure (expenditure or demand side), by measuring the total income (income side) and finally by measuring the total value of delivered production factors (factor side). This last option, however, is a mere theoretical possibility. The recent UN framework document that aims at an international harmonization of national accounting (System of National Accounts, SNA 1993) only lists the first three approaches to measure GDP (see Carson, 1995). Although the circulation models on which modern national accounts are based are much more sophisticated than our illustrative case, they give rise to the same three different possibilities to measure output.

However, national accounts do not necessarily contain all three types of aggregation. The German national accounts, compiled by the German Federal Statistical Office (DESTATIS), measure GDP only from the production-side as well as from the expenditure-side (see DESTATIS, 2002). The Bureau of Economic

[4] For the underlying theoretical model see Brümmerhoff (2002) who gives a detailed introduction to the German national accounts.

Analysis of the US Department of Commerce, in contrast, uses the expenditure-method (called "product side") and the income-method in order to establish the US national accounts (U.S. Bureau of Economic Analysis, 2002).

*Precision of GDP-data*

How precise is GDP-data? In order to answer this question, we have to compare different GDP-data for one and the same economy and year. This poses some difficulty because, in contrast to GDP forecasting in which several institutions are engaged, the national statistical agency more or less has a monopoly on GDP measurement. This is not due to some regulation but simply to the fact that it takes such a lot of work to establish the national accounts and to estimate the GDP.

Although there is but a single source of GDP-data, we are nonetheless able to evaluate the GDP-data's precision, namely by comparing the data that is (i) obtained according to the different ways in which the statistical offices actually calculate GDP, or (ii) released at different points in time. Let us start with the latter.

The quarterly as well as the annual GDP-data is subject to successive revisions in the years subsequent to its first release. Such revisions might be due to (i) corrections and completion of the data-basis or (ii) conceptual change in the system of national accounts. The first type of revisions occurs because national accounting faces the problem of large gaps in the basic statistics and therefore involves indirect methods of estimation. With the availability of new data, for example from a recent census, some extrapolations might be corrected. Additionally, some basic data such as the emergence and disappearance of new firms is only available with a certain time-lag and has to be completed in the course of a revision (Young, 1995, p. 428).

Changes in the system of national accounts such as the implementation of new calculation-methods, the adoption of a new price-basis for the calculation of *real* GDP, or the new development of methodological concepts, definitions and classifications necessitate a recalculation of former GDP estimates (see Essig & Hartmann, 1999). Indeed, the conceptual framework of the national accounts has been subject to constant change from the very beginning of its establishment after World War II.[5] This process is still ongoing: The SNA 1993 characterizes itself as "a stage in the evolution of national accounting".[6] It would, however, be a mistake to attribute these perpetual conceptual changes to the mere insufficiency or inadequacy of the old systems. No, the change rather stems from the need to adapt to new economic developments: the emergence of new products, new industries and

[5] For the younger history of national accounting see Kendrick (1995).

[6] Cited from Brümmerhoff (2002, p. XIV).

| | Data (a) | Conceptual (b) | Total (c) |
|---|---|---|---|
| 1991 | +1.5 | +1.4 | +3.0 |
| 1992 | +1.2 | +1.3 | +2.5 |
| 1993 | +1.0 | +1.3 | +2.3 |
| 1994 | +0.8 | +1.2 | +2.0 |
| 1995 | +1.2 | +1.1 | +2.4 |
| 1996 | +0.7 | +1.1 | +1.8 |
| 1997 | +0.2 | +1.2 | +1.4 |
| 1998 | −0.1 | +1.2 | +1.1 |

Table 6.1: The German 1999 revision: difference between former and revised GDP-data in percentage points due to changes in the underlying data-basis (a), conceptual change (b) and total (c). (Source: Essig & Hartmann, 1999)

even whole sectors, changes in the way people work, in the role of homework, the importance of the hidden economy, in the tax system, new financial assets and the need for better international comparability of GDP-data. All these are driving forces for conceptual change in the national accounts.

Now, how large are GDP revisions? The last comprehensive revision of the German national accounts took place in 1999 and consisted in the implementation of the European System of National Accounts (ESA) from 1995. Table 6.1 shows to which extent the annual GDP-data (in current euro) of the 1990s has been revised. It thereby distinguishes the revision due to change in the data-basis on the one hand and due to conceptual change on the other hand. The difference between original and revised data (c) must be considered as very important. It is of the same order as the annual GDP-growth and surpasses (in the early 1990s: largely) the mean absolute forecast errors that were reported in part 1. This is even true if only the revision due to data-correction is considered; the relatively small difference in 1997 and 1998 does not undermine this conclusion since the corresponding GDP-values will certainly be subjected to further revisions. However, table 6.1 seems to be hardly consistent with results by Weidmann (2002) who found the mean absolute revision (MAR) of the GDP-revisions in Germany to equal 0.57% (1965-2000) and even 0.54% (1983-2000). Whereas we calculate a MAR of more than 2% from table 6.1! Weidmann defines GDP-revision as the difference between current and advanced GDP estimates without giving any more detail or a source. It is quite possible that he did not consider comprehensive revisions such as the large 1999 revision but only annual revisions; or he evaluated the advanced and the final (not the latest available) official estimates of GDP. Anyway, even a MAR of 0.54% is indicating significant data-imprecision and hence explaining forecast failure.

Let us now have a look at the GDP-data and its revisions in the US. Young (1995) presents and discusses five studies examining revisions of quarterly US

GDP estimates and altogether covering the period from 1947 to 1991, where revision is defined as the difference between the initial and the latest available GDP-value in percentage points. The first official GDP-values were subjected to serious revisions as shown by a MAR of more than 3% for the first decade after World War II. However, GDP-data improved significantly in the 1960s insofar as revisions were stabilizing on a lower level, varying between 1 and 2%, where they prevailed until 1991. In accordance with this finding, Fixler & Grimm (2002), examining GDP revisions for the years 1983-2000, conclude that "[since] the early 1980s, the revisions to the annual rates of change — without regard to sign — in the quarterly estimates of current-dollar and real GDP have averaged somewhat more than 1 percentage point" (p. 9). To sum up the US evidence: The MAR of GDP estimates is pretty exactly of the same size as the MAE of GDP (growth) forecasts. There is significant imprecision in GDP-data that can at least partly explain the systematic forecast failure.

Having discussed GDP revisions, we will now focus our attention on the different measurements that yield GDP-data and compare their results. These are, for Germany, the expenditure-approach and the production-approach and, for the US, the expenditure-approach and the income-approach. However, we should be careful not to identify the difference between this data with the overall measurement imprecision. For data that stems from the two approaches is not independent because both the statistical agencies of Germany and the US implement various measures already in the course of the preparation of the respective estimates in order to minimize the discrepancy between the two results (see DESTATIS, 2002; Young, 1995). In addition, both types of measurement make use of the same statistical methods which may introduce further bias. The statistical discrepancy between the results obtained by different approaches has to be considered as a lower boundary of the total data imprecision.

Table 6.2 shows the mean average discrepancy between the US GDP-data obtained by expenditure-approach and the income-approach from 1978 to 1987. It accounts for different revisions enabling us to see whether the discrepancy is lower for revised than for current data — this would in fact indicate that revised data were more precise. Surprisingly, this is not the case! While the discrepancy is on average less for the first annual revision than for the preliminary estimate, it rises continuously in the course of the subsequent revisions to its maximal value of 0.91% for the latest available estimate. Ongoing revisions and corrections tend to increase the statistical discrepancy between the measurement methods.

In contrast to the US, the official German GDP estimate is a qualified average of the values obtained by the two different approaches. The Statistical Office actually considers the final integration of the two values only as a "last step"

| (Revised) Estimate | Discrepancy |
|---|---|
| Preliminary | 0.65 |
| Final | 0.64 |
| First annual | 0.55 |
| Second annual | 0.72 |
| Third annual | 0.76 |
| Latest available | 0.91 |

Table 6.2: Average size of the discrepancy between quarterly US GDP-values obtained by the expenditure- and the income-approach as a percent of GDP (1978-87). (Source: Young, 1995, p. 444)

(DESTATIS, 2002, p. 31) which highlights once more that the two measurements are not independent. Since only the integrated data is officially released, available evidence on the discrepancy is weaker than for the US. However, it is briefly mentioned in DESTATIS (2002) that the discrepancy in the final 1995 values was equal to roughly 0,8% of the GDP, suggesting that the German estimates are not more precise than the US estimates. Interestingly, the official, integrated estimate in 1995 was 0.7% higher than the value obtained by the production-approach and only 0.1% lower than the expenditure-data. This makes clear that the integration is not a mere averaging but involves further judgment of the underlying assumptions. These and further assumptions underpinning the whole measurement of the GDP are to be illuminated in the next subsection.

Obviously, our assessment of GDP-data imprecision together with the general discussions above yields quite a good explanation of macroeconomic forecast failure. In short: How could a GDP-forecast be more precise than the data it is based on?

*Assumptions underlying the preparation of GDP estimates*

Before discussing the assumptions underlying GDP estimates in some detail, let us listen to a few general remarks relating to the quality of GDP-data. Neubauer (1994) with a view to the underlying data-basis of GDP estimates stresses,

> By no means do all data sets stem from direct surveys that have been conducted for this very epistemic purpose. Much information has to be gained by a combination of data that was originally not produced for such a combination. A lot of data has to be produced by extremely creative estimation. Even today, one might still say with a mixture of respect and anxiety: National accounting requires a little bit of magic. (p. 61)

It is largely agreed in the literature that the underlying data-basis exhibits serious loopholes .[7] On the one hand, such gaps necessitate "creative" estimation, according to Strohm,

> Given the incompleteness of the statistical foundation, setting up a national account has to be considered as a creative process rather than as a calculation. (p. 685)

On the other hand, they increase the importance of subjective judgment of the employee in the statistical office (Brümmerhoff, 2002, p. 109). For Hamer (1970), arbitrariness is the inevitable result,

> Once the estimator has processed all information and comes up with an interval of possible results that he judges as equally credible, it is principally irrelevant and indeed more or less arbitrary which of these values is eventually published. (p. 83)

Hence, "nobody knows the 'true' value [. . .] of the GDP" (Brümmerhoff, 2002, p. 109). And Strohm, again:

> Nobody knows the true value of the GDP. The figures published by the German statistical office represent the 'best' figures as considered by those responsible for their calculation and publication. In principal it is always preliminary. (p. 684)

This corresponds with the appraisal of Stephen McNees who considers macroeconomic forecasting, the preparation, and the revision of estimates as a "continuous process that starts long before the period concerned, and continues long after" (Öller & Barot, 2000, p. 294). Neubauer (1994) draws our attention to a dangerous consequence of this similitude between forecasts and estimates,

> The user of the data always runs the risk of singling out regularities, relationships or structural stabilities that the statistical office has put in the data when estimating it because results for domains where no direct survey is available can only be obtained by comparative estimation or regression. (p. 62)

Hence, if GDP estimates are, like some GDP forecasts, derived from past data by statistical time series methods, forecasts and estimates necessarily correspond to each other to some extent. This suggests that we have actually underestimated forecast failure in part 1 and it explains why forecast errors are lower for preliminary than for final data (see p. 40 above).

---

7 As for instance by Strohm (1997, p. 684f.), Brümmerhoff (2002, p. 109), DESTATIS (2002) or U.S. Bureau of Economic Analysis (2002).

With these general remarks in mind and in line with the discussion in section 6.3, the first particular problem we will investigate is that of how to fill the gaps in the underlying meso data. Such gaps occur for instance because the basic statistics are established with very different frequency, ranging from monthly (e.g. VAT-statistics) to multi-annual (e.g. comprehensive census). Statistics that are established less frequently than once per quarter have to be extra- or interpolated in order to be considered in the GDP estimate. Such extra- or interpolations are typically performed by assuming a correlation between an indicator-variable and the variable to be extrapolated. Clearly, there can hardly be empirical evidence in favor of this correlation-assumption since data of the extrapolated variable is very scarce *by definition of the problem.* However, this method plays an important role in GDP measurement.[8]

The second way to fill the gaps in the basic statistics is to calculate the missing data indirectly. Two such examples from the German production-approach are: the agricultural production which is calculated from the area under crops and the mean yields, and, secondly, the apartment rents which are derived from total assets of apartments and mean rents per square-meter. A more complicated and interesting example, discussed by Brümmerhoff (2002), is the hidden economy. There is obviously no complete record or direct data of hidden and irregular economic activities. Thus, relationships to available indicators have to be presumed — but cannot be confirmed empirically. The available indicators are: the evolution of monetary circulation, the difference between official and 'normal' rate of employment, the supposed causes of irregular economic activity, and finally the difference between the GDP-values obtained by different measurement-approaches — which in fact means that the hidden economy is used to stuff the statistical discrepancy. A last example of indirectly inferred meso data from the US income-approach and closely related to the hidden economy is the adjustment for the misreporting of income and expenses and the nonfiling of taxes,

> Although BEA adjusts for misreporting and nonfiling, the adjustments are developed from information that is only available periodically — with the exception of corporate profits, the most recent information [was, in 1995,] for 1988 or earlier. Because little is known about how taxpayer misreporting and nonfiling may vary with respect to either changes in tax laws and regulations or business conditions, the annual and quarterly pattern, particularly since 1988, is based largely on judgment and is subject to considerable error. (Young, 1995, p. 430)

8 See DESTATIS (2002, p. 22) and U.S. Bureau of Economic Analysis (2002, p. 20).

DESTATIS (2002) tried to quantify to which extend extra- and interpolations or indirect methods are used to complete the data-basis. While roughly 30% of the basic data has to be inferred indirectly under the production-approach (p. 16), this is the case for almost all the data underlying the expenditure-approach (p. 28).

But even if basic data is available, it might not fit into the concepts of the national accounts and thus requires manipulation that involves additional, problematic assumptions. Three examples should serve as illustrations of this point which has generally been raised in the preceding section: the inclusion of non-market goods in the national accounts, seasonal adjustment and deflation.

The GDP measures the economy's output in (current or constant) prices. But not all the products are traded on markets and valued by a certain price. In order to ensure comparability, fictitious prices have to be assigned to such goods and services, as Brümmerhoff (2002) elucidates,

> In such cases, one has to specify shadow prices, i.e. prices that approximate the potential market price. Yet there is generally not only one single value. Instead, many alternative fictitious prices (depending on the presumed use) are possible each of which can rely on more or less doubtful assumptions. (p. 91)

This problem of fictitious prices is particularly present in the whole governmental sector. Krelle (1967), stressing the absence of any means of checking the underlying assumptions, even concludes that the inclusion of non-market products defamiliarizes the meaning of the basic economic concepts used in the national accounts,

> As a real market price rarely exists one has to rely on estimates which are in fact, because of their principal untestability, conventions. [...] One plugs in figures one does not know what to do with, and notions such as salary, profit etc. thus get a meaning they do not have in ordinary language use and that can only be grasped after the study of the applied calculation methods. (p. 185)

So, the inclusion of non-market activities in the national accounts is based on assumptions that cannot possibly be checked, thereby introducing a large amount of uncertainty into the whole measurement process.[9]

Seasonal adjustments, serving as our second illustration, seek to neutralize the impact of periodic events such as weather, holidays etc. on quarterly GDP estimates. This manipulation involves questionable assumptions that quantify the impact of periodic events. They are based on a combination of different methods

[9] For further discussion of this issue also see Levin (1995).

including ARIMA time-series models. Comparing two different seasonal adjustments, one of them being based on some additional years of data, Young (1995) was able to approximate the imprecision introduced into quarterly GDP estimates by seasonal adjustment to 0.8% — which is, again, roughly the size of our forecast errors.

As a final example, let us consider deflation, i.e. the neutralization of price change in order to estimate RGDP. According to Young (1995), this creates "some of the most difficult measurement problems associated with the national economic accounts" (p. 432). In order to deflate a transaction booked in the national accounts, both its quantity and its price have to be known. So, a first problem emerges if only the transaction's quantity is known. This is the problem of deflating non-market goods. It is solved by generally assuming that fictitious prices change proportionally to market-prices of similar products. In the inverse case, when a transaction is only evaluated in monetary terms and its quantity is unspecified (e.g. subsidies), global price-deflators are used to neutralize inflation's impact. And the index formula actually used to calculate global price deflators has a considerable impact on the real GDP estimate (Young, 1995, p. 446). However, even if both the quantity and the price of a certain transaction are known, a couple of problems remain as it is for instance necessary to "remove the element of the observed price change that is associated with a change in the characteristics of the product — that is, with a change in quality. For products that are undergoing rapid technological change, the information this separation requires may be substantial and difficult to acquire." (Young, 1995, p. 432)

The upshot so far is that the systematic failure of macroeconomic forecasts can at least be explained partly by the serious problems relating to the quality of GDP-data. Macroeconomic forecasts are indeed affected by the poor quality of their underlying data through all the different mechanisms we have identified in the general sections of this chapter.

## 6.5 Data-quality in climatology

Data of main climate variables as for example temperature and precipitation is of importance for climate forecasting both in its unaggregated meso form (for the initialization of GCMs) and in its globally aggregated macro form (for testing GCMs). As we have just seen, poor data-quality of GDP-data can at least partly explain forecast-failure in economics. In line with our comparative approach, we investigate in the following whether this is also true for climatology.

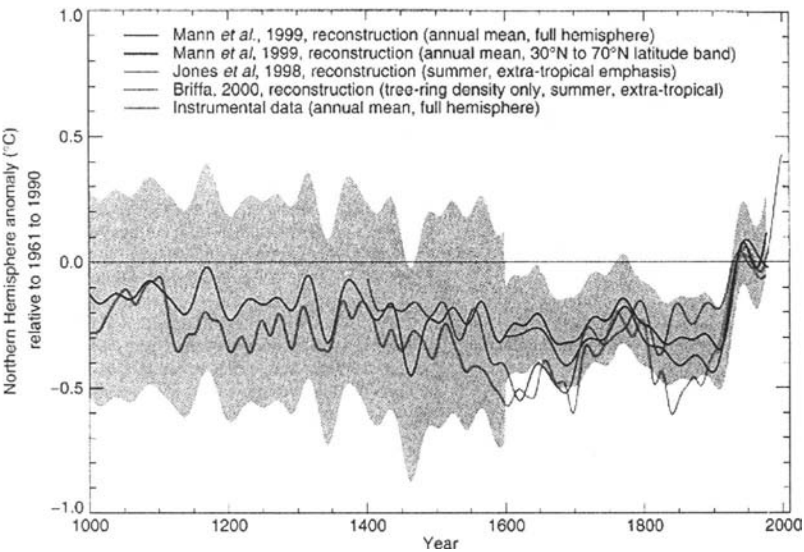



Figure 6.4: Three palaeoclimate temperature reconstruction of the past 1,000 and 600 years respectively. Besides plotting a more recent reconstruction as well as the instrumental data record, the diagram highlights the estimated standard error interval for one of the three reconstructions (shading). (Source: IPCC, 2001*b*, p. 134)

*Precision of climate-data*

The degree of imprecision of aggregated climate variables has already been indicated in figure 4.10 above. For in addition to the characteristic errors of 16 different climate models, it plots the correlation of alternative data, labeled (n) and (e), with the data the evaluation is based on. It is evident from the diagram that the statistical discrepancy between different climate-data is almost as high as the errors of model simulations!

Figure 6.4 compares three independent multi-proxy, i.e. based on a multitude of different indicators, temperature reconstructions of the last 1,000 years. One of the three time-series is shown with its uncertainty-interval (shaded) so that the comparison with the other two alternative series is facilitated. As the IPCC (2001*b*) notes, "significant differences between the three reconstructions are evident during the 17th and early 19th centuries where either [two alternative] series lie outside the estimated uncertainties in the [third] series" (p. 133). This is another strong evidence for poor precision of climate-data.

*Assumptions underlying the preparation of climate-data*

As we have seen in chapter 4, modern climate predictions are based on GCMs and demand a large amount of data both of high resolution and reaching back in time several hundreds, if not thousands of years. Clearly, available meso data does not meet these requirements: The instrumental observation of ocean and land surface weather variables did not start before the mid 19th century, upper air observations were not carried out before the 1950s (IPCC, 2001*b*, p. 96). And, relating to resolution, the density of observation stations is still very inhomogeneous. Paradoxically, the number of stations providing weather data is even declining! To make things worse, there are additional gaps in the data-records of the last 150 years that stem from wars and political change.

The closure of such temporal and spatial gaps in the records is frequently achieved through simple interpolation assuming that the data is sufficiently homogeneous. This procedure which relies on simplified assumptions is illustrated by the treatment of weather balloon data (IPCC, 2001*b*, p. 120) or the aggregation of sea surface temperatures (SST) and night marine air temperatures (NMAT) that we will discuss below.

But how can temperatures and other climate-data be obtained for periods before the instrumental observation started, e.g. for the last 1,000 years? Obviously, indirect methods of measurement that rely on theoretical assumptions in order to calculate climate variables from available data have to be applied. Indeed, climatologists have developed plenty of such indirect methods, described in the TAR (IPCC, 2001*b*, p. 130ff.), each relying on different indicators ("proxies"). Let us have a brief look at them while paying special attention to their underlying assumptions.

First of all, palaeoclimatic temperatures and precipitation can be calculated from *tree rings*, assuming that they relate to the rings' width and density. Hereby, non-climatic growth factors should ideally be taken into account, which is a major problem. The reconstruction of past temperatures from geochemical characteristics of *coral skeletons* as for instance from variations in trace elements or stable isotopes is quite similar to this method. As for tree rings, this exceeds knowledge on the exact impact of temperature on the corals' growth. Further indirect methods essentially consist in digging: *Ice cores* can provide several indicators as for example the fraction of melting ice or the rate of accumulation of precipitation. Insofar as annually laminated *sediments* in lakes and oceans are assumed to be influenced by climatic variations, summer temperatures and winter snowcover might be inferred. However, dating is a major problem for both these procedures. Yet another digging-method is even more problematic: That is *borehole* measurement. Oversimplifying assumptions concerning the geothermal properties of the earth near a borehole allow to relate the vertical temperature-profile in the borehole to the historical temperature-record. In addition, non-climate factors as land-use change, soil moisture, winter snowcover etc. etc. have to be considered and understood. Finally, *mountain glaciers moraines* might serve as a proxy, too. But it is very difficult to relate the extent of glaciers to regional climate change because of the important influence of local factors such as topography or ice accumulation.

Besides the questionable assumptions of all these methods, they are — without any exception — regionally biased! Thus, they just provide partial pictures of past temperatures and have to be combined for a spatially more complete record. In addition, regional bias makes the direct comparison of the different proxy-data difficult or impossible. We will discuss at the end of this section how this diverse meso data is integrated in spite of these problems.

Hitherto, we have primarily discussed how climate-data can be obtained for times and regions for which no instrumental observations are available. However, even instrumental climate-data has to be manipulated because of changes in instrumentation or measurement technique or insufficient shielding, for example. This point shall be illustrated by three examples: sea surface temperature, data of the upper atmosphere and satellite data.

Mean sea surface temperatures are calculated from more than 80 million single observations, primarily from the UK Main Marine Data Bank and the U.S. Comprehensive Ocean Atmosphere Data Set (see IPCC, 2001*b*, pp. 110f.). This meso data requires careful adjustment before it can be aggregated. SSTs are for example corrected for the use of uninsulated wooden and canvas buckets to collect water before the 1940s! From then on, better isolated buckets or ship-engine water has been used to perform the measurements. The case of night marine air temperatures is quite similar: NMATs had to be corrected for the progressive increase in the height of thermometer screens before 1930. In addition to this, they suffer a warm bias during World War II which has to be neutralized.

Pervasive changes in balloon instrumentation give rise to the necessity of correcting upper air data. Some of the temperature observations have to be corrected by as much as 3°C. However, according to the IPCC (2001*b*), it appears to be difficult to determine exactly when instrumentation has changed, as "alternative methods for identifying change points yield different trend estimates and [...] the analysis was hampered by the lack of complete documentation of instrumentation and data processing changes for many stations" (p. 119).

The need for data-manipulation does, however, not only arise for observations made with old-fashioned instrumentation. Satellites may certainly count as some of the most advanced, extravagant and costly instruments science ever had at its disposal. Yet satellite data has to be corrected for several effects, each correction involving assumptions for which there is more or less evidence. First of all, there are orbital and atmospheric transmission effects that have to be taken into account (IPCC, 2001*b*, p. 96). Furthermore, the non-linear calibration coefficients of one of the longest-lived weather satellite turned out to be erroneous (IPCC, 2001*b*, p. 120). That clearly gives rise to the need for corrections based on assumptions of the error-size. Additionally, satellite orbit decay introduces gradual cooling and thus gives rise to a cold bias of the data. Responses of the satellite's instruments often differ between the laboratory assessment and on-orbit performance, which causes more trouble. And, finally, there are diurnal effects induced by the east-west drift of the spacecraft. In sum, satellite data does not require less correction involving much less problematic assumptions than any other climate-data.

Hence, climatologist have very different kinds of data at their disposal, all corrected and adjusted for very different reasons and based on very different assumptions. How can the scientist be sure that this data-basis is self-consistent? The answer is this: "So-called data assimilation systems have been developed, which combine observation and their temporal and spatial statistics with model information to provide a coherent quantitative estimate in space and time of the climate system." (IPCC, 2001*b*, p. 96) These data assimilation systems do not only make sure that the different meso data match but also that the 'data' meets the requirements of high spatial and temporal resolution. In fact, the climate-data we have discussed so far only serves as an input to climate-models that "reanalyze" the data so that it satisfies the modeling requirements. Or, as the IPCC (2001*b*) puts it, "The principle of reanalysis is to use observations in the data assimilation scheme of a fixed global weather forecasting model to create a dynamically consistent set of historical atmospheric analysis" (p. 120). The IPCC (2001*b*, p. 474) admits that model evaluation (and, consequently, construction) is primarily based on reanalysis-data. All this seems to sum up to a prime example of fatal theory-ladenness: The reanalyzed climate-data relies on models whose evidence stems from the reanalyzed climate data! This circularity leads us to the conclusion that the assumptions underpinning the generation of 'climate-data' are, if anything, weakly confirmed at most. Climate data hence is of poor quality: it is imprecise, and it is generated on the basis of problematic assumptions.[10]

In the climate case, too, we can explain forecast-failure with poor data-quality. This does not only shed an interesting light on climate forecasting and the science of climatology as a whole but it also strengthens our initial explanation of macroeconomic forecast failure by reference to poor quality of GDP-data.

---

[10] Poor data-quality is also highlighted by the fact that there is still some scientific controversy about the data. Thus in response to a recent critique of their temperature reconstruction (see figure 6.4) by McIntyre & McKitrick (2003) Mann et al. (2004) had to admit errors in the original listing of time series that were used as proxies.

# Chapter 7

# Expectations and reflexiveness


## Summary

This chapter starts by reconstructing an argument from Oskar Morgenstern against the possibility of economic forecasting which explains economic forecast failure by pinpointing forecast reflexiveness. Subsequently, two lines of attack against this argument are presented. One of them consists in the proposal by Grunberg & Modigliani to construct self-fulfilling forecasts. The other one is based on the rational expectations hypothesis. However, it is shown that both of these counter-arguments either fail or give rise to new explanations of forecast failure. Thus, whether the critiques are appropriate or not, forecast failure can be explained with reference to expectations and reflexiveness.


## 7.1 Morgenstern's argument for the impossibility of macroeconomic prediction

Contrary to predictions in natural sciences, the publication of economic forecasts might alter the behavior of economic agents, i.e. the very object of the forecast. This brings about the well-known problem of reflexiveness, also known as "Oedipus effect of forecasting" (Bestuzhev-Lada, 1993).

> The phenomenon of reflexiveness is easily illustrated. An economist surveys farmers' resources and opportunities and the current price of wheat, and plugging this data into her theory, she predicts that there will be a surplus this fall and that the price will fall. This prediction, circulated via the news media, comes to the attention of farmers, who decide to switch to alternative crops because they now expect lower wheat prices. The results are a shortfall of wheat and high prices. Here we have an example of a suicidal prediction. The dissemination of a

> physical theory has no effect on its subject matter, but the dissemination of a social theory does. It is reflexive [...]. (Rosenberg, 1995, pp. 115f.)

Further examples are abundant in financial markets where public forecasts of future asset prices have immediate effects on trading behavior and consequently on prices. More generally, whenever a forecast is published as a warning, it is assumed to be reflexive and (hopefully) self-falsifying. Altogether, three cases must be distinguished. (i) The dissemination of the macroeconomic forecast has no effect on the predicted event, it is neutral. (ii) The dissemination of the forecast brings about another than the predicted outcome, it is self-falsifying. (iii) The dissemination of the forecast causes the prediction to become true, it is self-fulfilling. Bearing this distinction in mind, we will clarify in the following why reflexiveness poses a problem for forecasting by reconstructing an argument by Oskar Morgenstern against the possibility of macroeconomic prediction. After we have made all its premisses explicit, the subsequent sections will challenge the argument and give rise to further investigations into the role of expectations in economics.

Morgenstern's argument can be divided into three major parts. In its first part, he tries to show that there is a causal relation between the publication of a macroeconomic forecast and the predicted event. Taking this as a starting point for the second part, he then deduces that macroeconomic forecasts generally fail. The third part finally excludes the possibility that they are corrected.

Morgenstern (1928) argues as follows:

> We face an economic system in which the behavior of economic agents is based on the more or less exact knowledge of, and the more or less justified conjecture about the behavior, i.e. the sum of economic actions, of the relevant individuals. (p. 92)

In modern terminology: Economic agents base their individual behavior on expectations relating to the (aggregated) behavior of the other economic agents. As an example consider a firm which has to set up its investment plan. Its decisions will obviously depend on the expected future demand in its market, i.e. the aggregated behavior of its potential clients. Now, Morgenstern continues:

> Here, the absolute-prediction comes into play: An agency, which has full credibility, assesses the current state of the system for a time period that directly follows the current one. Thus, it predicts the sum of all economic actions. This prediction, however, is considered as correct by the economic agents. Therefore, a new datum enters into each individual prediction: the knowledge of the others' behavior. This new

> datum is so important that it affects every economic action. The economic agents will hence reorganize their dispositions — or generally: decisions. (p. 93)

In a sentence, Morgenstern argues that macroeconomic forecasts causally influence their object via the mechanism that relates individual expectations and behavior. Applied to our example: Since firms typically base their sales forecasts on macroeconomic GDP-forecasts, the latter's publication influences the firms investment decisions which in turn determine GDP. As a consequence of Morgenstern's reasoning, the above-mentioned possibility (i) that economic forecasts do not influence their object is ruled out. Strengthening the causal chain might also be the reason why Morgenstern additionally assumes that every agent believes in the published forecast. However, even a macroeconomic forecast which is not absolutely credible is a "datum" that enters into the individual's expectation-forming process. Therefore, we drop this assumption in the following reconstruction. Finally, it should be noted that the conclusion of the first part of the argument is reached thanks to the implicitly assumed transitivity of the causal relation.

As a next step, Morgenstern establishes that a forecast which causally influences its object does not become true. That is, possibility (iii) is ruled out, there are no self-fulfilling forecasts. He argues that even in the case of seemingly self-fulfilling forecasts as for example a credibly predicted boom in stock prices, the

> extent of the real change in any case is much larger than the predicted one. The duration of the real events cannot coincide with the predicted one. (p. 95)

Morgenstern seems to assume that a forecast inevitably fails if it does not consider each of the causally relevant factors[1], and that it cannot anticipate the effect of its own publication. This interpretation is confirmed since our author, in the third part of the argument, considers an ex post revision of the original forecast as the only way to account for the fact that there is a causal relationship between the forecast and its object,

> Hence, one has to consider the question whether a correction of the prediction is possible so that it finally comes true. When the first forecast

[1] This is obviously too strong an assumption. It is absolutely possible that forecasts based on time-series methods which do not take into account each of the causally relevant factors are correct. The falsity of this assumption is one of the reasons why Morgenstern's ambitious project — to prove the impossibility of macroeconomic forecasting a priori — is doomed to fail. Yet, we reconstruct his argument as an a posteriori explanation of macroeconomic forecast failure and in such a context, the assumption seems to be quite acceptable. For if we observe that a forecast has failed and discover that a relevant causal factor had been omitted, this generally counts as an explanation.

> is published, it is possible to add a second one. This second forecast would recognize the existence of the first one as a datum for the economy and would have to integrate it. (p. 97)

Now, such a revision, although differing substantially from the original forecast in terms of the methods involved in its preparation, would be a further macroeconomic forecast of the same object as the original one. Thus, a causal relationship holds and the revision, as soon as it is published, will not become true either. Hence follows that "the second prediction requires a third one, this a fourth one etc. *ad infinitum*" (p. 98). From this infinite regress, it is clear that a macroeconomic forecast cannot be corrected. Assuming that forecasting in the respective domain is impossible if a forecast does not become true and cannot be corrected, we finally reach Morgenstern's conclusion. The whole argument can now concisely be written as,

(1) A published macroeconomic forecast has a causal influence on the economic agents' expectations.
(2) The economic agents' expectations causally influence their behavior.
(3) The aggregated behavior of economic agents is the object of a macroeconomic forecast.
(4) The causal relation is transitive.
(5) *Thus:* There is a causal relationship between the publication of a macroeconomic forecast and its object.
(6) Unless a forecast considers each of the causally relevant factors ex ante, it will not become true.
(7) Macroeconomic forecasts cannot anticipate the effects of their dissemination.
(8) *Thus:* Macroeconomic forecasts will not become true.
(9) A revision of a macroeconomic forecast is a macroeconomic forecast.
(10) If there is no correct revision (or revision of a revision or ... ) of a macroeconomic forecast, the forecast cannot be corrected.
(11) *Thus:* Macroeconomic forecasts do not become true and cannot be corrected.
(12) If a forecast of a certain object neither becomes true nor can it be corrected, forecasting in this domain is not possible.
(13) *Thus:* Macroeconomic forecasting is not possible.

The first two parts of this argument can be challenged. Counter-arguments against the second part, notably against premiss (7) will be presented in the next section, while in the last section, we will primarily challenge premiss (1).

## 7.2 Constructing self-fulfilling forecasts

Premiss (7) was challenged in a paper by Grunberg & Modigliani (1954). In order to show that macroeconomic forecasts can in principle anticipate the effect of their dissemination, the authors make two assumptions. First of all, they assume that "private" forecasting is possible, i.e. as long as forecasts are not published, they become true. This is a reasonable assumption that allows us to investigate the isolated effects of forecast publication. Furthermore,

> once private prediction is assumed to be possible, the agents' reaction to a public prediction must also be regarded as knowable. For the assumption that private prediction is possible implies that it is possible to ascertain (a) how the agents' expectations are formed and (b) how the agents act in response to given expectations. (Grunberg & Modigliani, 1954, p. 466)

This conclusion, however, is erroneous. Assume for example that perfect private prediction using ARIMA-methods were possible. This does not necessitate that the expectation forming process of economic agents is understood. What Grunberg & Modigliani thought to be an implication of the possibility of private forecasting is in fact a further, second assumption: The publication-effect of a forecast is known. Or, in the framework of Morgenstern's argumentation, correct *private* revision of published forecasts is possible.

Now, the private, revised forecast clearly depends on the originally published forecast. From our assumption, the following function is known ex ante,

$$_{t-1}X_t^R = R(_{t-1}X_t^F),$$

with $_{t-1}X_t^F$ denoting the first, published forecast and $_{t-1}X_t^R$ its correct private revision. Grunberg & Modigliani call this relationship "reaction function". Since private revision is assumed to be possible, i.e. $_{t-1}X_t^R$ will be the correct forecast as long as it is not published, the question is simply whether a forecast $_{t-1}X_t^*$ exists with $_{t-1}X_t^* = R(_{t-1}X_t^*)$! Such a forecast $_{t-1}X_t^*$ would become true although it is published. Indeed, it would become true because it is published since the publication is not assumed to be neutral. $_{t-1}X_t^*$ is what we usually call a self-fulfilling forecast.

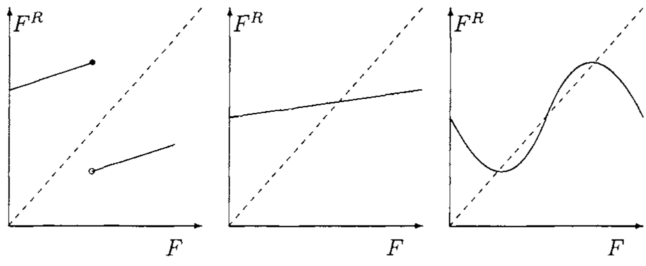



Figure 7.1: The graphs of three reaction functions with no, one and multiple self-fulfilling forecasts.

When do such self-fulfilling forecasts exist? The situation can be represented graphically by plotting the reaction function. If the reaction function crosses the diagonal, then a self-fulfilling forecast — which equals the X-coordinate of the intersection point — does exist. Figure 7.1 illustrates the three cases of no, one and multiple self-fulfilling forecasts. In this last case, the forecaster can, by choosing one of the self-fulfilling forecasts for publication, determine the actual outcome without deteriorating his prediction.

Formally, the existence of a self-fulfilling forecast is equivalent to the existence of a fixpoint of $R$. Applying *Brouwer's Fixpoint Theorem*, namely that boundedness and continuity are sufficient conditions for the existence of a fixpoint, Grunberg & Modigliani directly infer that at least one fixpoint exists (in $[k, K]$) whenever the reaction function has upper and lower bounds $K, k$ and is continuous on the interval $[k, K]$. Before discussing these results we shall illustrate them by a simple macroeconomic model.

Our discrete time-step model consists of three equations, two of them modeling aggregate demand and supply:

$$m_t + \bar{v} = p_t + y_t \tag{7.1}$$
$$y_t = \bar{y} + \beta(p_t - {}_{t-1}p_t^{\varepsilon}) \tag{7.2}$$
$${}_{t-1}p_t^{\varepsilon} = \Psi({}_{t-1}p_t^{F}), \quad \Psi' \neq 0 \tag{7.3}$$

where price level, $p_t$, real output, $y_t$, and the expectations at $t-1$ of the price level that will occur in the next period $t$, ${}_{t-1}p_t^{\varepsilon}$, are the endogenous variables. Accordingly, money supply, $m_t$, potential or full employment output, $\bar{y}$, the velocity of money, $\bar{v}$, and the published public forecast at $t-1$ of the price level that occurs in the next period $t$, ${}_{t-1}p_t^{F}$, are assumed to be given. Lower cases indicate that variables are represented by their respective logarithms. Obviously, equation (7.1) is nothing else but the simple equation of exchange that can be regarded as a short cut for aggregate demand modeling (see Romer, 1996, p. 245), whereas (7.2) is the so-called Lucas supply equation stating that the deviation of the actual from the potential output is proportional to the error of the public's price level expectation. It was derived for the first time by Robert Lucas from some simplifying microeconomic assumptions plus the assumption that *certainty-equivalence* holds, i.e. that under uncertainty economic agents behave as if the expected mean(s) were certain (see, again, Romer, 1996, pp. 246f.). Finally, equation (7.3) introduces the reflexiveness of the public forecast into the model assuming that the public price level expectations depend on the published forecast where $\Psi$ is a complicated function modeling the expectation forming process.

Now, we can calculate (privately) the price level that will occur at $t$ when all the exogenous variables, in particular the price level's forecast, are specified by solving the equation system. Substituting (7.2) into (7.1) yields:

$$m_t + \bar{v} = p_t + \bar{y} + \beta(p_t - {}_{t-1}p_t^{\varepsilon}). \tag{7.4}$$

Eliminating the expectation term with (7.3) and rearranging finally gives rise to

$$p_t = \frac{\beta}{1+\beta}\Psi({}_{t-1}p_t^F) + \underbrace{\frac{1}{1+\beta}(m_t + \bar{v} - \bar{y})}_{\text{const.}}. \tag{7.5}$$

This is the correct private forecast ${}_{t-1}p_t^R$, expressed as a function of the published forecast ${}_{t-1}p_t^F$; hence, (7.5) is nothing else but the reaction function in our model. Does it have a fixpoint? In order to answer this, the expectation forming function must somehow be specified. For the sake of simplicity, we assume that the public forecast is fully credible and $\Psi({}_{t-1}p_t^F) = {}_{t-1}p_t^F$. Thus, the fixpoint condition becomes

$${}_{t-1}p_t^* = \frac{\beta}{1+\beta}{}_{t-1}p_t^* + \frac{1}{1+\beta}(m_t + \bar{v} - \bar{y}),$$

which can be resolved to

$${}_{t-1}p_t^* = m_t + \bar{v} - \bar{y}.$$

Hence, if the public forecast were completely credible, forecasting the price level that occurs if actual equals potential output would be the only self-fulfilling forecast. On this condition, actual output would indeed equal potential output. However, it is an oversimplification to assume that the agents' expectations are identical to the public forecast whatever the public forecast predicts! As an alternative, we may try to apply the results from our discussion of the general case. If the reaction function (7.5) is bounded and continuous, the existence of a fixpoint — and, consequently, of a self-fulfilling forecast — is warranted. Is our reaction function bounded? Since the reaction function is linear in $\Psi(\cdot)$, this clearly depends on the properties of the expectation forming process. If for example $\Psi({}_{t-1}p_t^F) = {}_{t-1}p_t^F$ as assumed above, the reaction function is not bounded and the condition for applying Brouwer's theorem does not hold. What about the other condition? Is the reaction function at least continuous? As for boundedness, this depends on the expectation forming process. Do two expectations converge towards each other as their underlying public forecasts do? This sounds reasonable at first glance, but why should the expectation function not be a step function? Nobody can rule out this possibility. These problems make it virtually impossible to determine whether a self-fulfilling forecast exists — not to speak of its quantitative value.

I want to stress that these problems are not particular to my specific model but apply to the framework of Grunberg & Modigliani in general. Yet, these authors argue in favor of the general applicability of their arguments:

> To complete the argument, it must be shown that the (sufficient) conditions for the possibility of correct public prediction derived from Brouwer's theorem are likely to be fulfilled in the real world in which social events occur.
>
> (i) The variables of economic theory generally cannot increase or decrease without bounds. [...]
>
> (ii) Normally, the functions formulated in economic theory are conceived to be continuous. (p. 474)

If (i) were true, it would indeed follow that the reaction function is bounded. Grunberg & Modigliani refer to the physical boundaries of our economic system in order to prove (i). But this does not apply to nominal variables. There is no price level limit in times of hyper-inflation. And in addition, the reaction function might not even be bounded insofar as forecasts of real variables are concerned. The full-credibility case above is a clear counter-example. As to the second condition, the fact that (ii) functional relationships in economics are usually assumed to be continuous does not warrant at all that the reaction function is continuous. The reaction function necessarily involves a functional specification of the expectation forming process since this is the only mechanism through which published forecasts feed back on the forecasted event. Yet the problem is: We do not know what this function looks like. We even cannot rule out that it is not continuous. That is also admitted by Grunberg & Modigliani:

> The expectation function [...] offers particular methodological difficulties and is actually different from the other functions encountered in economic theory. Expectations are themselves predictions, ranging from the elaborate scientific forecast of the large business enterprise to primitive guesses and dark hunches. While the other economic laws

> assert what people will do, given (1) motivational postulates and (2) certain expectations, the expectation function asserts what people will predict, given certain past and current observations. In the predictive model, it predicts what the agents will predict. The familiar motivational postulates of economics are here of no use. Thus for some time to come the expectation function will, at best, appear in the form of broad statistical generalizations. (p. 471)

Do Grunberg & Modigliani not undermine their whole argument with this confession? They did not think so, because the fact that the expectation function $\Psi(\cdot)$ will hardly ever be specified "is not a difficulty of public prediction in particular, but rather of social theory and prediction in general". This, however, is another false conclusion that follows from our authors' belief that private prediction implies the possibility of private revision. But as we have seen, ignorance of the expectation function is not necessarily a problem for private macroeconomic forecasting.

We may conclude that Morgenstern's argument is still delivering a good explanation for macroeconomic forecast failure — not only because it is so difficult to account for the publication effect but also because this is simply not done in the course of actual forecast preparation. We do not specify and include reaction functions in the models used for macroeconomic forecasting. In terms of the reconstruction of Morgenstern's argument, this can be reflected by simply changing "cannot" to "do not" in premiss (7).

## 7.3 Forecast-neutrality under rational expectations

Premiss (1) of Morgenstern's argument stated the non-neutrality of public forecasts. It had been justified by the fact that a public forecast enters as a new datum into the expectation forming process of the economic agents. Now, assume that the economic agents already knew the true prediction, consequently, its publication would not have any effect on their expectations. Assume furthermore that they cannot be fooled (through the publication of false forecasts), then public forecasts would be completely neutral (case (i) from the very beginning of this chapter) and economic forecasting would not be reflexive.

As implausible this reasoning might sound at first glance, it has been one of the central hypotheses of macroeconomic theory in the last 30 years: the so-called rational expectations hypothesis (REH). REH was originally proposed by John F. Muth in 1961; he defines rational expectations as

> essentially the same as the predictions of the relevant economic theory. (Muth, 1961, p. 316)

More formally, Muth specifies:

> The hypothesis can be rephrased a little more precisely as follows: that expectations of firms (or, more generally, the subjective probability distribution of outcomes) tend to be distributed, for the same information set, about the prediction of the theory (or the "objective" probability distributions of outcomes). (p. 316)

Using the notation of Sheffrin (1996) with $I_t$ denoting all available information at time $t$, Muth's hypothesis can be put as

$$\begin{aligned} \text{subjective expectations} &= {}_{t-1}X_t^{\varepsilon} = \mathrm{E}[X_t | I_{t-1}] \\ &= \text{conditional expectations} \end{aligned} \tag{7.6}$$

Although REH was not only proposed but also applied by Muth, it is notably due to the work of Robert Lucas that REH has become a pillar of modern macroeconomic theory.

Clearly, REH undermines premiss (1), and "a 'public prediction' [...] will have no substantial effect on the operation of the economic system" as Muth (1961, p. 316) realized. In the framework of Grunberg & Modigliani (1954), REH implies that the reaction graph is a horizontal line. According to the REH, it is not the correct public forecast but the agent's expectations which are self-fulfilling.

In order to save Morgenstern's argument, REH must be rejected. But on which grounds? A strategy which seems to suggest itself is to point out that economic agents do not have rational expectations, i.e. REH is accused to be unrealistic. Yet this, I maintain, would be a fallacious strategy. REH is most appropriately interpreted as a principle telling us how to construct macroeconomic models. Here, Nancy Cartwright's theory of science can be helpful, again. In a nutshell, Cartwright (1999) distinguishes three levels of empirical science: theory, model and experiment. Theory consists of very abstract principles that tell scientists how to construct specific models for specific contexts. These models, in turn, represent such contexts and allow to perform experiments. The point is that, according to Cartwright, the theoretical principles do not represent anything on the experimental level. They only function as a guide for model-construction. This, I propose, is also the best view on REH. Specifically, as will be exemplified below, REH is a receipt for how to eliminate the subjective expectations term in a formal economic model, namely, by substituting it by the model's conditional expectations. If this view is appropriate, REH may not be interpreted realistically and it cannot be refuted by pointing out its 'unrealism' as Frydman & Phelps (1983) and Pesaran (1987) for instance do.

The thesis that REH may not simply be rejected by observing how economic agents form their expectations is supported by the fact that the meaning of "rational expectations" varies from model to model and from empirical application to application. This may even give rise to scientific controversies as the Barro-Small quarrel on whether REH implies that agents should have anticipated the effects of the Vietnam-war budget or not (see Sheffrin, 1996, p. 54).

However, REH does not entirely escape empirical testing since the assumptions and implications of a REH-model can be compared with observations. Thus, a REH-model can be rejected because its empirical implications turn out to be false. It therefore seems to be an absolutely sound strategy to save Morgenstern's argument by pointing out (a) that there is the phenomenon of reflexiveness as many examples of self-falsifying and self-fulfilling forecasts show and (b) that any REH-model is empirically inadequate with regard to such cases.[2] At this point, however, we will follow another route and investigate the implications of rational expectations for macroeconomic forecasting. It will become evident that while REH undermines the reflexiveness explanation of forecast failure, it gives, at the same time, rise to other limits of economic prediction. As REH is an abstract theoretical principle, its implications on forecasting also depend on the model to which it is applied. Whether REH allows for forecasting or not has therefore to be answered on the model level and thus, three different applications of REH will be discussed right now.

*1st application: modeling the supply side*

The first application of REH we will discuss is a variation of the model introduced in the last section. It comprehends the demand and supply equations (7.1) and (7.2). In contrast to the above model, money supply $m_t$ is an endogenous variable following a money supply rule $R$ with random shocks $e_t$,

$$\begin{aligned} m_t &= R(y_{t-1}, y_{t-2}, \ldots, p_{t-1}, p_{t-2}, \ldots) + e_t \\ &= R(\mathbf{y}_{t-1}, \mathbf{p}_{t-1}) + e_t \end{aligned} \tag{7.7}$$

where bold letters stand for the record of past values, and $\mathrm{E}[e_t|I_{t-1}] = 0$ for all $t$. Again, we assume that the economy is in equilibrium and thus may substitute (7.2) into (7.1) in order to obtain our well-known equation,

$$m_t + \bar{v} = p_t + \bar{y} + \beta(p_t - {}_{t-1}p_t^{\varepsilon}). \tag{7.8}$$

[2] Chapter 10 presents some evidence relating to the empirical adequacy of REH-models.

As a next step, we take conditional expectations relative to the information available at $t-1$ on both sides. Since mathematical expectations are linear, this yields

$$\begin{aligned} \mathrm{E}[m_t|I_{t-1}] + \mathrm{E}[\bar{v}|I_{t-1}] &= \mathrm{E}[p_t|I_{t-1}] + E[\bar{y}|I_{t-1}] \\ &\quad + \beta\mathrm{E}[p_t|I_{t-1}] - \beta\mathrm{E}[{}_{t-1}p_t^{\varepsilon}|I_{t-1}]. \end{aligned}$$

Because $\mathrm{E}(\cdot)$ is the operator of the model-intern expectations, i.e. $\mathrm{E}(x_t|I_{t-1})$ denotes the future value of $x_t$ as predicted by the model at $t-1$, this reduces to

$$\begin{aligned} R(\mathbf{y}_{t-1}, \mathbf{p}_{t-1}) + \bar{v} &= \mathrm{E}[p_t|I_{t-1}] + \bar{y} \\ &\quad + \beta\mathrm{E}[p_t|I_{t-1}] - \beta_{t-1}p_t^{\varepsilon}. \end{aligned} \tag{7.9}$$

Now, the REH enters the scene. Specializing (7.6) to $p_t$ it follows that $\mathrm{E}[p_t|I_{t-1}] = {}_{t-1}p_t^{\varepsilon}$ and we see that the last two terms of (7.9) are equal. Replacing conditional by subjective expectations and rearranging yields

$${}_{t-1}p_t^{\varepsilon} = R(\mathbf{y}_{t-1}, \mathbf{p}_{t-1}) + \bar{v} - \bar{y}. \tag{7.10}$$

Next, we substitute (7.10) and (7.7) into (7.8) which, after simplification, gives rise to

$$p_t = R(\mathbf{y}_{t-1}, \mathbf{p}_{t-1}) + \bar{v} - \bar{y} + \frac{e_t}{1+\beta}. \tag{7.11}$$

Subtracting (7.10) and (7.11) we find that the difference between realized and expected price level is $e_t/(1+\beta)$ and, finally, the Lucas supply equation (7.2) can be written as

$$y_t = \bar{y} + \frac{\beta}{1+\beta}e_t. \tag{7.12}$$

This equation states that deviations from potential output, i.e. its trend, are random and cannot be predicted!

*2nd application: modeling the demand side*

The second application of the REH takes the life-cycle hypothesis as a starting point, i.e. the assumption that households maximize the expected summed utility over their life cycle. This subjectively expected life-cycle utility of an individual with a live-span $[0 \dots T]$ is, at the beginning of her life,

$$ {}_0U^{\varepsilon} \;=\; {}_0\Big(\sum_{t=0}^{T} \frac{1}{(1+\delta)^t} u(c_t)\Big)^{\varepsilon} \tag{7.13} $$

where the instantaneous utility, $u()$, is a function of the households consumption at time $t$, $c_t$; and the rate of subjective time preference, $\delta$, quantifies the fraction of utility loss when a fixed amount of goods is consumed — ceteris paribus — in period $t+1$ instead of in $t$.

Equation (7.13), however, is not accessible for further analysis. The reason for this is that we do not know anything about the subjective expectations function $(\cdot)^{\varepsilon}$. Here, REH offers resort. According to REH we may replace the subjective expectations term with the mathematical expectations conditional to the information $I_0$. This allows for the following transformations

$$ \begin{aligned} \mathrm{E}[U|I_0] \;&=\; \mathrm{E}\Big[\sum_{t=0}^{T} \frac{1}{(1+\delta)^t} u(c_t)|I_0\Big] \\ &=\; \sum_{t=0}^{T} \frac{1}{(1+\delta)^t} \mathrm{E}[u(c_t)|I_0]. \end{aligned} \tag{7.14} $$

Assuming perfect capital markets, i.e. that households can borrow and save for a given interest rate $r$, and not forgetting that households have to fulfill their budget constraint, Hall (1978) was able to derive that the expected marginal utility, $u'()$, of the next period's consumption is equal to the current marginal utility times a certain constant,

$$ \mathrm{E}[u'(c_{t+1})|I_t] = \frac{1+\delta}{1+r} u'(c_t) \tag{7.15} $$

This is the so-called Euler-equation.

Here comes a semi-formal proof for it. Assume for the sake of argument that a household, $A$, takes consumption decisions so that equality did not hold. If $\mathrm{E}[u'(c_{t-1})|I_t] > (1+\delta)/(1+r)u'(c_t)$, consider an individual household, $B$, who is identical to $A$ except of saving an infinitesimal fraction $\triangle c$ at time $t$ in order to consume the saved assets plus the paid interest rate at $t+1$. Let the summed utility of periods $t$ and $t+1$ for the two households be $U^A$ and $U^B$ respectively, thus, via Taylor expansion,

$$ \begin{aligned} \mathrm{E}[U^B|I_t] \;&=\; u(c_t - \triangle c) + \mathrm{E}\Big[\frac{1}{1+\delta} u(c_{t+1} + (1+r)\triangle c)\Big| I_t\Big] \\ &\approx\; u(c_t) - u'(c_t)\triangle c \end{aligned} $$

$$
\begin{aligned}
& +\mathrm{E}\Big[\frac{1}{1+\delta}u(c_{t+1}) + u'(c_{t+1})(1+r)\triangle c)\Big| I_t\Big] \\
= \;& u(c_t) + \mathrm{E}\Big[\frac{1}{1+\delta}u(c_{t+1})\Big| I_t\Big] \\
& \underbrace{-u'(c_t)\triangle c + \frac{1+r}{1+\delta}\mathrm{E}[u'(c_{t+1})|I_t]\,\triangle c}_{>u'(c_t),\text{ as assumed}} \\
> \;& u(c_t) + \mathrm{E}\Big[\frac{1}{1+\delta}u(c_{t+1})\Big| I_t\Big] - u'(c_t)\triangle c + u'(c_t)\triangle c \\
= \;& \mathrm{E}[U^A|I_t].
\end{aligned}
$$

Consequently, household $A$ did not maximize the expected life-cycle utility in contradiction to the utility maximizing assumption. If $\mathrm{E}[u'(c_{t+1})|I_t] < (1+\delta)/(1+r)u'(c_t)$, a contradiction can be derived similarly by assuming that household $B$ borrows instead of saves consumption $\triangle c$.

What are the implications of the Euler-equation? As Hall (1978) states, (7.15) implies that

> [no] information available in period $t$ apart from the level of consumption, $c_t$, helps predict future consumption, $c_{t+1}$, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods $t$ or earlier are irrelevant, once $c_t$ is known. (p. 974)

In other words, current consumption is a sufficient statistic with regard to future consumption.[3] Hall implicitly assumes that expectations are rational and that $I_t$ consists of every available information at time $t$. In other words, that economic agents do not miss any kind of information when forming their expectations. Since $u'(c_{t+1}) = \mathrm{E}[u'(c_{t+1})|I_t] + e_{t+1}$ with $\mathrm{E}[e_{t+1}|I_t] = 0$, marginal utility follows a truly unpredictable random walk with trend, i.e.

$$u'(c_{t+1}) = \frac{1+\delta}{1+r}u'(c_t) + e_{t+1}. \tag{7.16}$$

Furthermore, if the utility function were quadratic and marginal utility were thus linear in consumption, $u'(c_t) = ac_t + b$, then substitution,

$$ac_{t+1} + b = \frac{1+\delta}{1+r}ac_t + b + e_{t+1},$$

and simplification,

$$c_{t+1} = \frac{1+\delta}{1+r}c_t + \frac{e_{t+1}}{a}, \tag{7.17}$$

[3] A random variable $X$ is a sufficient statistic with regard to $Y$ for information $I$ if and only if $\mathrm{P}(Y|X) = \mathrm{P}(Y|X, I)$.

show that consumption would follow a random walk, too. Hall finally concludes that

> [under] the pure life cycle–permanent income hypothesis, a forecast of future consumption obtained by extrapolating today's level by the historical trend is impossible to improve. The results [...] have the strong implication that beyond the next few quarters consumption should be treated as an exogenous variable. (p. 986)

*3rd application: modeling financial markets*

In our final example, REH is applied to financial markets where it gives rise to the efficient market hypothesis (EMH). According to the definition of Malkiel (1992),

> a capital market is said to be efficient if it fully and correctly reflects all relevant information $[I_t]$ in determining security prices.

Formally, in an efficient market, the current price is a sufficient statistic for the information $I_t$ with respect to future prices. In strict analogy to Hall's result, it follows that price fluctuations are random and cannot be forecast on the basis of $I_t$.

How does EMH follow from REH? To see this in a semi-formal way, consider a single asset whose price at time $t$ shall be denoted by $P_t$. Assume furthermore that there is no inflation, that traders who try to maximize profits can borrow for and save at interest rate $r$, and that the market mechanism operates. On these conditions, the traders' asset-price expectations are equal to the current price plus the interest it would yield if paid into an account,

$$ {}_{t-1}P_t^{\varepsilon} \;=\; (1+r)P_{t-1}. \tag{7.18}$$

If this equation did not hold, traders would borrow money and buy assets or sell assets in order to maximize profit and the market mechanism would restore equality.

By applying REH (7.18) becomes

$$\mathrm{E}(P_t|I_{t-1}) = (1+r)P_{t-1},$$

and given the fact that $P_t = \mathrm{E}(P_t|I_{t-1}) + e_t$, this yields

$$P_t \;=\; (1+r)P_{t-1} + e_t. \tag{7.19}$$

Equation (7.19) states that actual prices reflect all information relevant for future prices and available to traders. Prices follow a random walk with trend. Thus, we have 'derived' the EMH from REH under very restrictive assumptions (including profit maximization of identical agents and operating market mechanism). A much more sophisticated application of REH by Grossman (1978) shows that the identical agents assumptions (or, to be precise, universally available information) is not necessary to derive EMH from REH. In Grossman's model, each trader has her private source of information. Assuming (i) that the joint distribution of this private information and the security returns is normal and (ii) that the aggregated demand for securities is decreasing in its price, he derived, at a first step, that current prices are a sufficient statistic if and only if prices are a sufficient statistic in an "artificial economy" where each trader had access to all information. Secondly, he proved that actual prices are indeed a sufficient statistic in this "artificial economy". However, Grossman also demonstrated that freely available private information as well as zero transaction costs are necessary to deduce EMH from REH.

Hitherto, we have rather loosely talked about the information set $I_t$ to which expectations are conditional. Traditionally, three empirical specifications of this information set and, consequently, the EMH, are distinguished. The weak EMH assumes that $I_t$ consists of the historic record of security prices. Accordingly, forecasts based on current prices cannot be improved by considering the past record. Secondly, the semi-strong version of the EMH assumes that $I_t$ is identical to all publicly available information. Finally, according to the strong EMH, $I_t$ compromises any information known to at least one market participant. This strong version corresponds to Grossman's results but as he has shown, it crucially depends on the fact that insider information can be obtained at no cost. This suggests that the strong version might empirically be inadequate, which is the case, indeed (see Campbell et al., 1996). At the same time, however, there seems to be empirical support for the semi-strong and the weak EMH (Malkiel, 1992). The finding from chapter 2 that a random walk model can hardly be outperformed by other forecast methods is, of course, a further confirmation. Or, the other way around, the preceding reasonings explain those observations. And the mentioned study of Clarida et al. (2001) now appears in another light: Even if a newly developed model delivers more accurate exchange rate forecasts than the random walk benchmark, this superiority will disappear as soon as this result is made public because it is instantaneously exploited by market participants.

*Do random walks explain forecast failure?*

The random walk results we have obtained throughout this section crucially depend on the assumption that there (already) is some kind of uncertainty. In our first application of REH, there was uncertainty with regard to future money supply, in the second application, consumers were uncertain about future consumption levels (or, in the original model of Hall (1978), about future income) and, finally, in the financial markets application, traders did not know future prices with certainty. But was it not this lack of uncertainty and unpredictability that we tried to explain originally? Isn't the explanation of forecast failure by REH circular?

Admittedly, the implications of REH do not *ultimately* explain why economic processes are unpredictable. The main message can rather be put like this: If there is unpredictability of some kind in the system — and that seems very much to be the case —, the best we can do is to use naïve rules for forecasting. Simple extrapolations cannot be systematically outperformed by any forecast method because any information is already reflected in today's system state. Thus, REH seems to explain at least one of our major observations from part 1.

Let us finally listen to what the principle advocate of REH says about its impact on macroeconomic forecasting! According to *Business Week* quoted from Sheffrin (1996, p. 159),

> Lucas says he is "pessimistic" that the rational expectations theory can ever be used to develop mathematically quantifiable forecasts and that short-term forecasts, at any rate, are essentially extrapolations of current conditions. "And for that purpose," he says, "rational expectations doesn't have much to add."

# Chapter 8

# Sensitive dependence on initial conditions


## Summary

Sensitive dependence on initial conditions can explain forecast failure. This chapter develops a "routes to chaos"-approach which shows that all empirically adequate models of the economy as well as the climate exhibit sensitive dependence on initial conditions. As a first step, that approach identifies certain minimal models, namely the coupled oscillator and the cellular automata, which sensitively depend on initial conditions. Next, they are shown to model some characteristic features of both the economy as well as the climate under their every-day description. As the minimal models sensitively depend on initial conditions, so does any adequate model of the economy and our climate system, too.


## 8.1 Sensitive dependence on initial conditions and forecasting

*Explaining forecast failure with sensitive dependence on initial conditions*

A forecast's performance depends, as we have seen in chapter 6, on the predicting data's quality, in particular its precision. Imprecise data increases forecast failure. Yet, forecasts may be more or less robust to data imprecision, and the predicting model actually determines the degree to which forecast performance is affected by poor data quality. Now, if a forecast derived from a certain predictive model sensitively depends on the predicting data — so sensitively that data which does not agree in infinitely many decimals gives rise to completely different forecasts — it is merely this sensitive dependence on initial conditions (SDIC) which explains forecast failure; for no variables can be measured with an infinite precision.

To derive a more formal characterization of SDIC, we have to introduce some concepts of the theory of non-linear dynamics. Take a formal model consisting of a set of partial differential equations as a starting point. These equations determine the future evolution of the system for every possible set of initial states, i.e. for every point in the phase diagram.[1] Thus, they induce a function, the *flow*, that maps the phase space onto itself for some given time-step $\triangle t$. Now, an *attractor* is defined as a subset of the phase space that is invariant under the flow. Generally spoken, an attractor of a system is a set of system-states that, as a whole, is not affected by the system's dynamic. If a system has reached a state contained in an attractor, all its future system-states are included in that attractor, too. Simple examples of attractors are: fixed points (equilibrium positions), closed curves, or tori (oscillations). However, some attractors do not belong to any of these types, they are 'strange'. A dynamic system possesses a strange attractor if and only if it is sensitively depending on initial conditions (see Lorenz, 1993, p. 170). Such systems are also called "chaotic".

More precisely, a dynamic system exhibits SDIC if the deviation between two different time-paths increases exponentially. Expressing this idea more formally will enable us to derive a principal limit of forecasting in terms of a maximum forecast horizon. Thus, if the random variable $\eta_t$ represents the difference between two time-paths at time $t$ in general and $\eta_0$ represents the difference between the initial states (at $t = 0$), then the mean growth rate of the initial deviation is approximated by,

$$E(\frac{\eta_t}{\eta_0}) \approx e^{\lambda_+ t} \tag{8.1}$$

where $\lambda_+$ is a system-specific parameter called the greatest positive Lyapunov-exponent (Schreiber & Kantz, 1996, p. 47). As equation (8.1) shows, errors in the initial conditions grow exponentially if the system possesses a strange attractor.

This said, a principal limit of predictability in the sense of a maximum horizon of predictability can be derived.[2] Let the random variable $\delta_0$ represent the deviation between initial system states, i.e. the imprecision of data. Assuming that prediction becomes senseless for forecast horizons greater than $\tau$ if the forecast volatility due to data imprecision has grown to the size of the volatility of the predicted variable, $X$, itself, i.e. $\mathrm{V}(\delta_\tau) \geq \mathrm{V}(X)$, it follows that prediction is restricted to horizons

$$\tau < \frac{1}{2\lambda_+} \ln \Big( \mathrm{V}(X)/\mathrm{V}(\delta_0) \Big).$$

The derivation in detailed steps is,

$$\begin{aligned}
\mathrm{V}(\delta_\tau) &< \mathrm{V}(X) \\
\mathrm{V}(\delta_0 \cdot e^{\lambda_+ \tau}) &< \mathrm{V}(X) \\
e^{2\lambda_+ \tau} \cdot \mathrm{V}(\delta_0) &< \mathrm{V}(X) \\
2\lambda_+ \tau + \ln(\mathrm{V}(\delta_0)) &< \ln(\mathrm{V}(X)) \\
\tau &< \frac{1}{2\lambda_+} \ln \Big( \mathrm{V}(X)/\mathrm{V}(\delta_0) \Big).
\end{aligned}$$

[1] Look back at section 5.4 for an illustration.

[2] See also Kadtke & Kravtsov (1996, p. 8), Anosov & Butkoskii (1996, p. 68) and Kravtsov (1993, pp. 176ff.).

*Detecting sensitive dependence on initial conditions*

Having seen that SDIC can explain forecast failure, the question arises how to detect it. How do we figure out that the predictive models suffer from SDIC? Since macroeconomic as well as climate predictions rely on very different methods and models, and since, as we saw in part 1, the predictive failure does not depend on the underlying method, it would not be a completely satisfying explanation of macroeconomic forecast failure to demonstrate that a certain type of model exhibits SDIC. Such a *single model approach* is insufficient. On the other hand, it is simply impossible to show that the economy 'as such' or the climate 'as such' sensitively depends on initial conditions and disturbances. For SDIC is a property of mathematical descriptions of systems so that, in order to reveal whether the economy or the climate 'as such' exhibit SDIC, we would have to know its 'true' model — a condition, if meaningful at all, we do not fulfill. The *metaphysical approach* fails, too. The dilemma between showing too little (one single model exhibits SDIC) and too much (the system 'as such' exhibits SDIC) might be solved by a more empirical, *statistical approach.* Its central idea is that certain characteristics of the data imply that only non-linear models which sensitively depend on initial conditions can describe the corresponding system in an empirically adequate way. Such tests typically measure the fractal dimension of the embedded time-series, but are characterized by several shortcomings (see Cawley et al., 1996). First of all, the intended purpose of such statistical methods is to discriminate between 'deterministic chaos' and 'randomness', that is to detect any underlying structure in the data-set at all. Such structures might, of course, occur in random data, too; or, on the other hand, they might be destroyed by small random noise. Furthermore, not only does one have to estimate the 'true' dimension of the system (facing the same problems as the metaphysical approach), but this method also works for low dimensional systems only. Altogether, there is no definitive statistical test for SDIC.

In contrast to statistical tests for SDIC which have been criticized for completely ignoring additional qualitative knowledge of the system, the approach applied in this chapter is solely based on such qualitative system-characteristics. Yet, such a qualitative approach is not unproblematic, either, as Lorenz (1993) points out,

> [It] should be noted, however, that chaos in continuous-time dynamical systems cannot be established via general and simultaneously simple characteristics of these systems [...]. During the last decades, a variety of higher dimensional systems belonging to different families has been investigated proving the presence of a strange attractor, but it is not

> always clear whether the diverse examples possess common (possibly hidden) structural properties. (pp. 173f.)

What can a qualitative test look like then? Reaffirming that there is no set of disjunctively necessary and conjunctively sufficient conditions for SDIC in terms of a system's structural characteristics, Nicolis & Prigogine (1998) make the following proposal,

> [The] mechanism for the emergence of chaotic attractors cannot be cast in a universal, normal form [...]. Rather, by combining some general considerations with the experience acquired from extensive numerical computations, we arrive at regularities that allow us to identify certain particular routes to chaos. It is difficult to assert whether this is an inevitability due to the very nature of chaos, or a temporary drawback due to the complexity of the problem. (p. 129)

Such a "routes to chaos"-approach comprehends the following steps.[3] First of all, one identifies a minimal mathematical model, $\mathcal{M}_{\mathrm{min}}$, that realizes a simple, general structure and exhibits SDIC (which can be shown analytically or by numerical simulations). The second step consists in highlighting certain essential structural, qualitative features of the system under its usual, every-day description, $\mathcal{D}$, which are captured by the minimal mathematical model. Finally, one infers that any empirically adequate model, $\mathcal{M}$, of the system sensitively depends on initial conditions, too. This inference, as illustrated by figure 8.1, is warranted by the following argument,

(1) $\mathcal{M}_{\mathrm{min}}$ models some essential structural, qualitative features, $\mathcal{D}_{\mathrm{S}} \subset \mathcal{D}$, of the system $\mathcal{D}$, i.e. $\mathcal{M}_{\mathrm{min}}$ is homomorphic to $\mathcal{D}_{\mathrm{S}}$.
(2) If some model $\mathcal{M}$ is structurally adequate of $\mathcal{D}$, then $\mathcal{D}_{\mathrm{S}}$ is homomorphic to $\mathcal{M}$.
(3) A model $\mathcal{M}$ is empirically adequate of a system only if it is structurally adequate.
(4) *Thus:* $\mathcal{M}_{\mathrm{min}}$ is homomorphic to $\mathcal{M}$ if $\mathcal{M}$ is an empirically adequate model of $\mathcal{D}$.
(5) SDIC is a structural property of systems, that is if some $\mathcal{S}_1$ is homomorphic to some $\mathcal{S}_2$ and $\mathcal{S}_1$ exhibits SDIC, so does $\mathcal{S}_2$, too.
(6) $\mathcal{M}_{\mathrm{min}}$ exhibits SDIC.
(7) *Thus:* Any model of $\mathcal{D}$ is either empirically inadequate or sensitively depends on initial conditions.

[3] I reconstruct the approach in analogy to Nicolis & Prigogine (1998, pp. 93ff.).

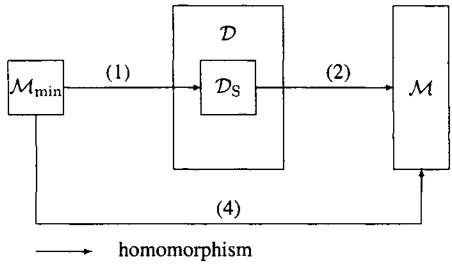



Figure 8.1: Illustration of the argument underlying the "routes to chaos"-approach.

As different structural features of a system can be highlighted, and different minimal models may be considered, there are possibly alternative arguments which show that a system exhibits SDIC, and a whole network of routes to chaos might emerge. The following sections explore two alternative routes: While one of them is based on the coupled oscillators-model, the other one uses cellular automata as minimal model.

## 8.2 The coupled oscillators route

*Introducing the minimal model*

The qualitative features highlighted under the coupled oscillators approach are feedback processes inherent to the dynamic system. The importance of feedbacks for the emergence of strange attractors is widely acknowledged and stressed.[4] Simply spoken, "feedback" refers to processes which comprehend causal loops and are consequently either self-dampening (negative feedback) or self-increasing (positive feedback). This said, what does a formal model look like that includes a feedback? Let us consider a dynamic model in its most general form,

$$\begin{aligned} \dot{x_1} &= f_1(x_1 \dots x_n, \lambda) \\ &\vdots \\ \dot{x_n} &= f_n(x_1 \dots x_n, \lambda), \end{aligned} \tag{8.2}$$

and assume that it correctly grasps the causal structure of the modeled system.[5] Then, the system exhibits a feedback if and only if there is a set of state variables $(x_{k_1} \dots x_{k_l})$ with $\partial f_{k_i}(x_1 \dots x_n)/\partial x_{k_{(i+1)}} \neq 0$ for all $i = 1, \dots, l-1$ and $\partial f_{k_l}(x_1 \dots x_n)/\partial x_{k_1} \neq 0$. Models that satisfy this condition typically give rise to

[4] See for instance Nicolis & Prigogine (1998) and Ormerod (2000).

[5] That is the processes represented by the variables on the left hand side are caused by the phenomena represented by the variables on the right hand side.

various kinds of oscillations, i.e. they possess closed orbits as attractors for some parameter values. The one-dimensional, simplified Duffing oscillator,

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= y_1 - \lambda y_1^3 - y_2, \end{aligned}$$

is an example for a nonlinear system with one positive and three negative feedback-loops. Principally, $2d$ state variables are necessary to model an oscillator with $d$ degrees of freedom through a set of first-order differential equations.

Since the solution of a dynamical model, namely its attractors, depends on the parameter values, new types of attractors might occur if the parameters vary. This phenomenon is referred to as *bifurcation*. A special type of bifurcation is the metamorphose of a fixed point into a periodic orbit attractor due to the increase of a parameter: it is called *Hopf-bifurcation*. Likewise, a close attractor may bifurcate into a 2-dimensional torus due to further parameter increase (secondary Hopf-bifurcation), which may in turn become a 3-dimensional torus etc. However, the number of successive Hopf-bifurcations is limited by the system's dimensionality. As Seydel (1994, p. 97) argues, Hopf-bifurcations are "abundant" in non-linear systems since the conditions for their existence are sufficiently weak and not affected by perturbations.

Hitherto, we have focused on individual oscillators and isolated systems with inherent feedbacks only. But as its name already indicates, the coupled oscillator route assumes certain kinds of interference among the individual systems. The coupling of the oscillators corresponds — on the model side — to the fact that state-variables of other oscillators enter into the differential equation of an oscillator via so-called coupling terms. If we assume $x_i \in \mathbb{R}^{2d}$, the equations (8.2) are the general form of $n$ coupled $d$-dimensional oscillators. The whole system shall have a fixed point attractor for low values of the parameter $\lambda$, then a Hopf-bifurcation may occur by increasing the parameter (see Lorenz, 1993, p. 178). If the dimension of the system is sufficiently high, subsequent Hopf-bifurcations into more-dimensional tori occur through further parameter increase. In case of three successive Hopf-bifurcations, the important *Theorem of Newhouse, Ruelle and Takens* can be applied (Newhouse et al., 1978). It states that if the three frequencies characterizing the three-dimensional oscillation on the torus are incommensurable, small perturbations will transform it into a strange attractor. Or, in other words, "when the third frequency is about to appear, simultaneously a strange attractor arises, since a three frequency flow is destroyed by certain small perturbations. Bifurcation into a 3-torus is not generic" (Seydel, 1994, p. 339). This implies that a sufficiently complex coupled-oscillators system sensitively depends on initial conditions.

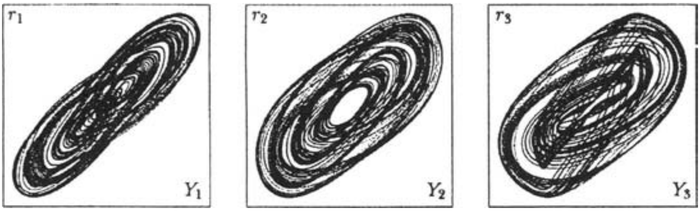



Figure 8.2: The attractors of three coupled Kaldor models projected into the $Y$-$r$ plane. (Source: Lorenz, 1993, p. 182)

*Applying the minimal model*

In order to show that the economy as well as the climate fall under this coupled oscillators approach, we have to show that they incorporate interdependent, positive and negative feedback processes.

Conceiving the economy as being composed of linked subsystems that incorporate feedbacks seems to be a fundamental characteristic of the way we use to describe the economy. The underpinning reason is the presence of information processing beings in the economy as we describe it, that is: human agents. Human agents introduce feedbacks because their behavior, which is based on information about economic processes on the one hand, determines these very processes on the other hand. If we acknowledge that the agents' supply- and demand-related decisions are influenced by corresponding market-prices, every market — no matter whether it is a financial, consumer, business-to-business or whatever kind of market — might serve as an example of this feedback mechanism. Over and above that, consumer decisions, to cite an example from Ormerod (2000), (i) are frequently influenced by and, at the same time, (ii) do influence the product's success. On the macroeconomic level, saving- and investment-decisions are conceived as being determined by aggregated macro-variables as GDP. Here, feedbacks occur because GDP is largely influenced by total savings and investment.

The vast majority of these subsystems have to be modeled as nonlinear oscillators. And somehow, they are all linked with each other. Thus, the preconditions of the coupled oscillator approach are met so that it can come into operation: The economic system possesses a strange attractor.

Yet the application of non-linear, coupled oscillators models in economics not merely are a theoretical possibility as the following example taken from Lorenz (1993) shows. The Kaldor model,

$$\begin{aligned} \dot{Y} &= \alpha(I(Y,K) - S(Y,K)) \\ \dot{K} &= I(Y,K) - \delta K, \end{aligned}$$

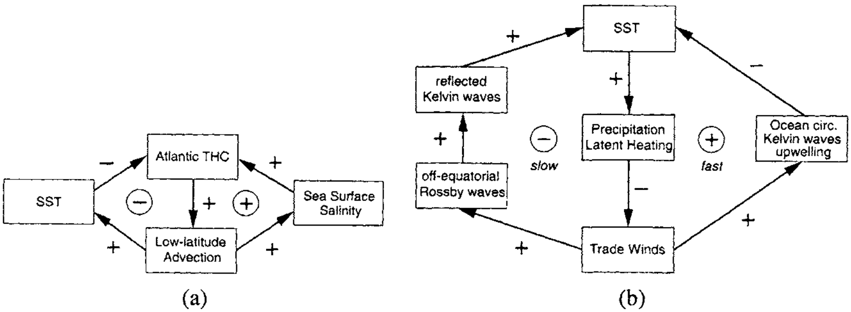



Figure 8.3: Simplified feedback loops involving the thermohaline circulation (a) and the El Niño-Southern Oscillation (b). (Source: IPCC, 2001*b*)

with nonlinear investment function $I$, linear savings function $S$, output $Y$, capital stock $K$ and $\alpha, \delta > 0$, is a macroeconomic IS-LM model. Having shown that the model is an oscillator (Lorenz, 1993, p. 47), Lorenz considers three economies obeying the Kaldor model that are coupled via international trade. Technically, money supply is introduced into the model as a third state variable whose time derivative depends on exports and imports $\dot{M} = Ex(\bar{Y}) - Im(Y)$ where $\bar{Y}$ is the 'rest of the world-production' and $Ex(\ldots)$, consequently, the coupling term. Numerical simulation of the model reveals the existence of a strange attractor for certain parameters. Figure 8.2 is a projection of the attractors of the three coupled economies into the output-interest rate plane.

This said and done, is there a similar route to chaos in climatology? First of all, feedbacks are a fundamental characteristic of the climate system, too. They occur at different spatiotemporal scales, in the atmosphere, the ocean, in land-surface processes as well as in the global, coupled system. A whole chapter of the IPCC's Third Assessment Report is devoted to this topic (IPCC, 2001*b*). Some of the most important feedback processes associated with global warming are the water vapor and cloud feedbacks: The greenhouse gas effect causes rising temperatures which affects the amount and distribution of water vapor in the atmosphere. Since water vapor is a potent greenhouse gas itself, a feedback loop occurs. In the same time, clouds coverage is related to atmospheric water vapor. Because clouds have an impact on the outgoing long-wave radiation as well as the incoming solar radiation — both of which influence global temperatures —, a second feedback loop emerges.

Figure 8.3 visualizes two further feedback processes on a global level that involve well-known climate oscillations, namely the thermohaline circulation and the El Niño-Southern Oscillation. As all the different climate oscillators are linked with each other, the coupled oscillators-scenario applies. We thus may conclude with the IPCC (2001*b*, p. 422) that the "Earth's atmosphere-ocean dynamic is chaotic: Its evolution is sensitive to small perturbations in initial conditions."

## 8.3 The cellular automata route

*Introducing the minimal model*

While the oscillator route focused on feedback processes, the cellular automata approach applies to systems that are composed of a possibly large but limited number of subsystems whose development is determined locally by adjacent subsystems. Such systems can be modeled by a cellular automata (CA). What is a CA? It is a system that consists of a finite number of sites between which a neighborhood relation is defined. Each site can embrace finitely many different states. Their evolution takes place in discrete time steps and is determined by a recursive, universal rule in function of their neighbor sites' states. Formally, a CA can be defined as (i) a graph $G(V, E)$ inducing a neighborhood relation between the sites $V$, plus (ii) a function $F_t : V \rightarrow S$ with $t = 0, 1, 2 \ldots$ and $|S| \in \mathbb{N}$ which is defined for all $v \in V, t > 0$ recursively by (iii) a rule $R$ with $F(v, t) = R(F(w, t-1) \mid \forall w : \{w, v\} \in E)$ and some initial conditions. Usually but not necessarily, the sites of a CA are a regular lattice in a $n$-dimensional space.

Let us first of all consider very simple CAs whose sites can exhibit two different states only and are arranged in a row so that each site has two neighbors — call them minimal CAs (MCA). The universal rule that governs the evolution of the sites in a MCA has to determine the state (black or white) of a site $v$ as a function of the former states of the site $v$ and its two neighbors. Thus, the recursive rule distinguishes $2^3$ different input states. There are consequently 256 different rules (as there are 256 different functions $R : \{1 \ldots 8\} \rightarrow \{0, 1\}$) and, correspondingly, 256 different MCAs. Stephen Wolfram was the first to investigate these CAs systematically (Wolfram, 1986). The evolution of MCAs as well as of other one-dimensional CAs can nicely be visualized in a diagram of horizontally stacked rows, with the $i$th row depicting the state of the automata at $t = i$ (see figure 8.4). Applying this visualization, Wolfram was able to distinguish four mutually exclusive, collectively exhausting classes of CAs,

> **(1)** Class 1. These cellular automata evolve after a finite number of time steps from almost all initial states to a unique homogeneous state (all the cites have the same value). The set of exceptional initial configurations which behave differently is of measure zero when the number of cells $N$ goes to infinity. [...] From the point of view of

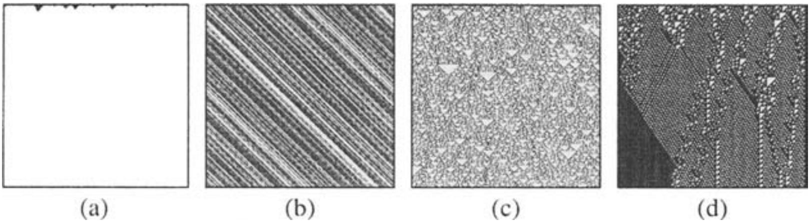


Figure 8.4: Examples for the four Wolfram classes: (a) Class 1, (b) Class 2, (c) Class 3, (d) Class 4. (Source: Chopard & Droz, 1998, p. 23)

> dynamical systems, these automata evolve towards a simple *limit point* in the phase space.
>
> **(2)** Class 2. A pattern of separated periodic regions is produced from almost all initial states. The simple structures generated are either stable or periodic with small periods. [...] Here again, some particular initial states (set of measure zero) can lead to unbounded growth. The evolution of these automata is analogous to the evolution of some continuous dynamical systems to *limit cycles*.
>
> **(3)** Class 3. These cellular automata evolve from almost all initial states to chaotic, aperiodic patterns. [...] Small changes in the initial conditions almost always lead to increasingly large changes in the later stages. The evolution of these automata is analogous to the evolution of some continuous dynamical systems to strange attractors.
>
> **(4)** Class 4. For these cellular automata, persistent complex structures are formed for a large class of initial states. The behavior of such cellular automata can generally be determined only by explicit simulation of their time evolution. (Chopard & Droz, 1998, pp. 22f.)

Figure 8.4 shows an example for each of these four classes.

The classes 3 and 4 are obviously of special interest for us. Their sensitive dependence on initial conditions is illustrated in figure 8.5. However, are these two classes generic? That clearly depends on the CA's design. As Gerhardt & Schuster (1995, p. 87) report, classes 1 and 2 are predominant for MCAs. Yet if CAs with an enlarged neighborhood are considered (extended from two to six), only one fifth of all possible automata belong to these 'robust' classes, i.e. the vast majority depends sensitively on initial conditions. A similar effect can be observed if CAs of higher dimensions are examined. According to Gerhardt & Schuster (1995, p. 90), the "evolution of two-dimensional [binary] automata is largely governed by Class 3." So, even relatively simple CAs exhibit SDIC and an increase of the CA's complexity seems to make SDIC even more likely.

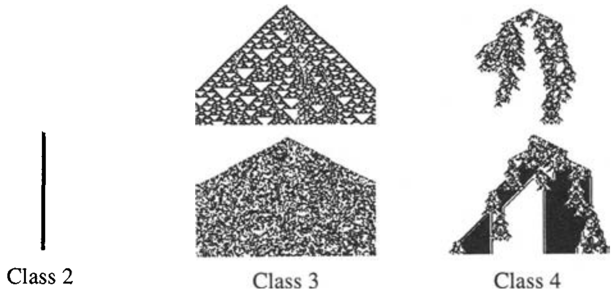



Figure 8.5: Difference diagrams of two exemplary cellular automata for each of the Wolfram Classes 2-4. A difference diagram highlights those sites whose states differ if only one initial state is altered. (Source: Gerhardt & Schuster, 1995, p. 85)

*Applying the minimal model*

The cellular automata route leads to SDIC in economics as well as in climatology because both the economy and the climate are conceived as consisting of separated components whose evolution is determined 'locally' by nearby components.

One key concept that allows us to apply the CA model to the economy is that of interacting agents, extensively discussed by Paul Ormerod in his *Butterfly Economics* (Ormerod, 2000). It is an almost trivial observation indeed that the behavior of economic agents is influenced by the behavior of other agents. Let us nonetheless consider some examples: Whether or not a consumer will for instance buy a certain product (e.g. watch the latest movie) frequently depends on whether his friends did so or not. A second example relating to stock markets is Keynes' "castle-in-the-air theory" of financial markets (see Malkiel, 1999, pp. 31f.). According to this theory, investors are largely influenced by their colleagues' decisions and behave in accordance with psychological principles that typically lead to herd behavior. This causes 'bubbles' whose regular occurrence is documented in the second chapter of Malkiel's *A Random Walk down Wallstreet* which is equally amusing and instructive. The output-side of the economy, as another example, exhibits a CA-structure, too. Output is produced by a large number of different firms who interact with other companies — for instance in their supply-chain — resulting in spill-over effects. Regional clustering is one observable effect.[6] On the macro scale, international trade may serve as an example, since the national economies interact with their neighbors. Globalization can be modeled as an extension of the individual neighborhoods. A very abstract feature of the economic system, namely the presence or absence of cooperation between agents, can be conceived as CA, too, as is demonstrated by several models exposed by Ger-

6 A concrete model implementing this idea is discussed by Ormerod (2000, pp. 170ff.).

| *cellular automata* | *site* | *state* | *neighborhood* |
|---|---|---|---|
| consumer market | consumer | product (not) acquired | friends |
| financial markets | investor | asset bought or sold | colleagues, gurus |
| output-side | firm | production | business sector |
| international trade | country | imports and exports | trading partners |
| cooperation | agent | cooperate/ non-cooperate | social environment |

Table 8.1: CA-structures in the economic system.

hardt & Schuster (1995, pp. 235ff.). These models may also serve as examples of more complex CAs with a large set of possible states and a complicated rule of evolution. Table 8.1 summarizes the examples of this paragraph underlining the structural similarity between the economy and CAs.

Do such structural similarities also apply to the climate? Recall the overall design of general circulation models of the climate system as represented in figure 4.9. GCMs dissect the atmosphere, the ocean and the land-surface by a 3-dimensional lattice. Relevant variables and vectors are attributed to the subspaces and a set of differential equations is used to determine the temporal evolution of the system locally, i.e. as a function of adjacent subspaces. This basic design looks very much like a CA. However, there are some differences. First of all, there are infinitely many different states a site of a GCM can embody while CAs are characterized by a limited number of different states. In this respect, GCMs resemble lattice Boltzmann models: These are modified CAs which allow for infinitely many states representing probabilities and which are used to simulate fluids and gases (see Chopard & Droz, 1998, p. 123). Secondly, a GCM evolves in continuous time and not in discrete time steps as CAs do. This said, we should recall that GCMs are solved numerically by computer programs. This has very interesting consequences, since solving a system of differential equations numerically involves reducing it to a discrete time system. In addition, our binary computers are not able to represent real numbers with infinite accuracy. Each variable and vector of the GCM is represented by a finite string of bits that can be in finitely many different states only! Thus, insofar as a GCM is run on a computer, it perfectly exhibits the structure of a cellular automata.

Let us conclude: As there are no long-range forces, the climate system can be conceived as consisting of many different sub-components whose evolution is determined locally. It can hence be modeled by a CA — with a large number of different possible states in order to ensure sufficient accuracy. If we drop the prerequisite of fast computability, there is no reason not to increase the number of possible states as much as appropriate. As we have seen, this potentially increases the automata's sensitivity to initial conditions.

Finally, it is interesting to notice that the two independent routes to chaos we have pursued might even be connected. Comparing the examples in Nicolis & Prigogine (1998) with those given by Gerhardt & Schuster (1995) and Chopard & Droz (1998), it is absolutely astonishing to see how many processes can be modeled by each of these two approaches in the same time: diffusion processes, autocatalytic chemical reactions as for example the Belousov-Zhabotinsky-reaction, predator-prey ecosystems and many others. The fact that all of these systems exhibit SDIC seems to strengthen the claim that the routes actually lead us to where they are supposed to.

## 8.4 Justified objections

The last section of this chapter is devoted to two objections which might be raised against the claim that the failure of macroeconomic and climate forecasts can be explained by the corresponding system's SDIC.

The first objection pinpoints that the occurrence of SDIC in our minimal models depends on the parameter values. In case of the coupled oscillators approach, a strange attractor appears only for sufficiently high values of the corresponding parameter $\lambda$. Similarly, with a parameter $\lambda$ defined for a binary CA as the probability that a state will be black in the next round, Class 3 automata are roughly characterized by $\lambda > 0,3$ (Gerhardt & Schuster, 1995, pp. 92f.). Are 'the parameters' of the economic and climate system sufficiently high so that strange attractors actually occur? Admittedly, we do not know. We do not even know what 'the parameters' are.[7] Consequently, we have not shown that the climate or the economy *are* actually (e.g. today) sensitively depending on initial conditions. Concluding that every empirically adequate model exhibits SDIC does not imply that infinitely small measurement errors *always* lead to complete forecast failure. All we have shown is that this might be the case! Yet, as a partial explanation of already observed forecast failure — and that is what we aim at —, that should be sufficient.[8]

The second objection criticizes a seemingly implicit assumption of our reasoning, i.e. that SDIC of parts implies SDIC of the whole. Skepticism vis-à-vis this principle seems to be expressed by Kadtke & Kravtsov (1996), saying

[7] Thus, during my visiting stay at the Potsdam Institute for Climate Impact Research a climate scientist confessed in a discussion that climatologists have not yet identified all control parameters that determine the attractors of the ocean dynamic.

[8] See also footnote 1 of chapter 7.

> that it is still not clear whether non-linear dynamics will be generally applicable to the real world, at least not in its present simplicity. All of these methods to some degree assume that the physical system in question have low-dimensional dynamical generators or representations; as yet, it is not known how realistic an assumption this may be for the Universe as a whole. (p. 19)

However, in my interpretation, Kadtke & Kravtsov refer to special forecasting methods based in non-linear methods or to strategies for detecting chaotic signals in noisy time-series. It is indisputable that these methods cannot be applied to large systems as the earth climate or even the universe as a whole as long as they assume low dimension. But we have not done so.

Now what about the principle 'SDIC in parts implies SDIC in the whole'? It is surely as questionable as the application of specific methods developed for low-dimensional systems to high-dimensional ones. The reason for this is that the behavior of very many chaotic subsystems may appear simple and regular if aggregated onto a larger scale. The trajectories of $H_2O$ molecules in a wave-reservoir might sensitively depend on initial conditions while the reservoir with the waves as a whole forms a robust, highly predictable system. In terms of the abstract argument of section 8.1, premiss (5) thus seems to be false. Nevertheless, this does not seriously undermine our reasoning since we do not need to rely on the strong "SDIC in parts implies SDIC in the whole"-principle. As the subsystems we have chosen, e.g. the consumption- and output-side of the economy, international trade or the gulf-stream, are of the same scale as the whole system itself, their effect on the whole system is not simply averaged out. With the amendment to premiss (5) that $\mathcal{S}_1$ is not merely mapped onto a tiny subset of $\mathcal{S}_2$ (and corresponding adjustments of the further premisses), the argument can hence be saved.

Not as an objection but rather as a reminder, let me finally highlight a basic assumption of our reasoning, namely that the way we talk about the economy and the climate is the benchmark of empirical adequacy. If the every day description is no longer shared, the newly described system does not necessarily exhibit SDIC. With some fantasy one might invent a description of the climate system under which no feedbacks and no oscillations occur. Under such a description, the THC or El-Niño could, for instance, just *seem* to be oscillations as the blinking lights at a Ferris wheel only *seem* to be an oscillator. It is for these reasons that I stated in the introduction that our explanations hold conditional to the language we speak, and that this inquiry is written from an internalist perspective.

# Chapter 9

# Experiment and simulation


## Summary

Simulation, according to this chapter's main result, will never improve the predictive performance neither of economics nor of climatology by compensating the lack of experiments. It is argued that there are principally two approaches to simulation: a conservative and a constructive one, and that both approaches cannot successfully be applied to the sciences this inquiry deals with. Last but not least, the methodological framework is used to unmask the absurdity of the ensemble-prediction method in climatology.


## 9.1 The promise

On 16th July 1945, the United States set off a nuclear pile in the desert of Alamogordo for the first time in the history of mankind. This has been the beginning of a series of scientific experiments involving unseen destructive powers. More than 2,000 nuclear tests have been performed until today. The Lawrence Livermore National Laboratory (LLNL), founded in 1952, has been one of the major institutes promoting this military research.[1] However, after the fall of the Berlin Wall, George H. W. Bush declared a moratorium on nuclear testing in 1992. Only three years later, Bill Clinton announced a stockpile stewardship program that aims at refurbishing, manufacturing and replacing nuclear weapons without relying on nuclear tests. Instead, the principal question "When does a weapon fail?" was supposed to be answered by *simulation*. According to Heller (2001), some of the most powerful super-computers are operating at LLNL to accomplish this demanding task.

This case is remarkable in so far as computer-simulation seems to successfully replace experimentation. Simulated nuclear tests seem to extend our phenomeno-

[1] For the laboratory's history see Middleton (2002).

logical knowledge, e.g. they tell us that a certain bomb will not explode unless this-and-this component is replaced. This is the promise of simulation social scientists hope to be fulfilled: being a substitute for experimentation. As we all know, this would be crucial as there is no experimentation involving the objects of social science, namely societies, due to practical as well as moral reasons. Practically, we lack the apparatus to sufficiently shield and arrange societies in a way necessary to perform (replicable) experiments. Even more important, this would be morally unacceptable. This holds, in analogy, for our atmosphere and climate system. In the following, I will use a rather broad *definition of simulation* that is not restricted to computer based simulations alone: The research method of substituting a system, called "simulating system", for the (original) system of interest, called "simulated system", and investigating the simulating system with the aim of obtaining insights into the simulated one is called simulation.

What does all this have to do with forecasting? Our investigation in chapter 4 has suggested that experimentation is an important condition for high forecast quality: First of all, predictive performance is an implication of the very idea of replicability. Secondly, predictive progress in depth as well as in scope stems from improved experimental skills. And finally, we can add that experiments play an important role in the context of discovery, i.e. the discovery of structures, capacities, regularities, etc., too, and hence contribute to the construction of (predictive) models. The question thus is: Will simulation (ever) be a satisfying substitute for the experimental method in economics and climatology and thereby improve their predictive performance? With regard to the forecast evaluations of part 1, it must also be explained why this has not yet been the case.

Jacobsen & Bronson (1997) give an optimistic outlook:

> The difficulties of testing sociological theories are well-known. [...] Simulations with computerized models can overcome or avoid most of these difficulties. [...] The constraints on experimentation disappear when dealing with computer models. (p. 97)

I, however, want to defend the opposite claim: Simulation cannot neutralize the lack of experiments and hence will not improve the predictive performance by replacing experimentation.

## 9.2 Reliability of simulations

Obviously, no simulation in the sense of the above definition will give rise to correct conclusions about the simulated system. Assume, for instance, that we were aiming at predicting the influence of a high-energetic electro-magnetic shock on

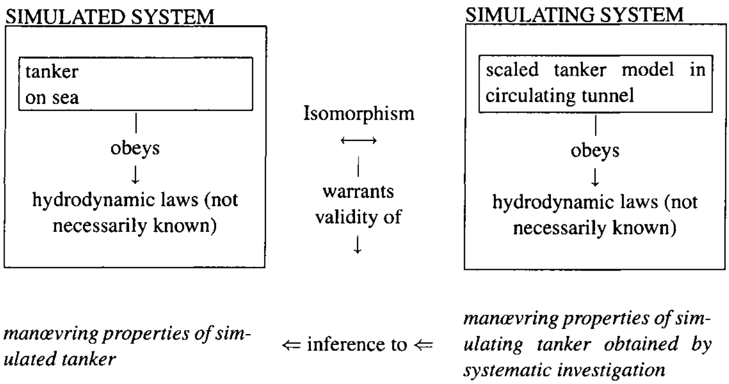



Figure 9.1: The manœuvring properties of a ship can be discovered by systematic investigation of a simulating system that is obtained according to the conservative approach. The fact that the simulating system is a close copy of the simulated one ensures structural similarity, i.e. that both systems are isomorphic, which in turn warrants the inference from characteristics of the simulating system to those of the simulated one.

the functioning of a nuclear bomb. In order to do so, we substitute a conventional firework that is studied systematically for the original system. The result of this investigation is: The functioning is not influenced by electro-magnetic shocks. This, however, is not the case with (at least some) nuclear weapons. Something is wrong with this simulation. The transfer of the results which were obtained by investigating the simulating system to the simulated one is not truth-preserving. A simulation that, contrary to our example, satisfies this condition will subsequently be called a *reliable* simulation. In our context, reliability is the central property of a simulation. If simulations are not reliable, they cannot replace experimentation as they do not give rise to insights into — and in particular categorical or conditional forecasts of — the simulated system, as experiments do. Unreliable simulations do not improve forecast performance.

What makes a simulation a reliable one? I argue that structural similarity (in all relevant aspects) between the simulating and the simulated system is a necessary condition: Assume such similarity did not hold, then both systems would behave differently under some interventions and evolve unequally. The inference from the simulating to the simulated system would not be truth-preserving. The absurd simulation of a nuclear weapon by a firework is for instance characterized by important structural differences between the two systems. In the light of these findings, it becomes clear that the inference involved in simulations is nothing but an inference from analogy; and such an inference is only valid if a sufficiently strong analogy, i.e. structural similarity, really holds.

We can generally distinguish two distinct ways of producing structurally sim-

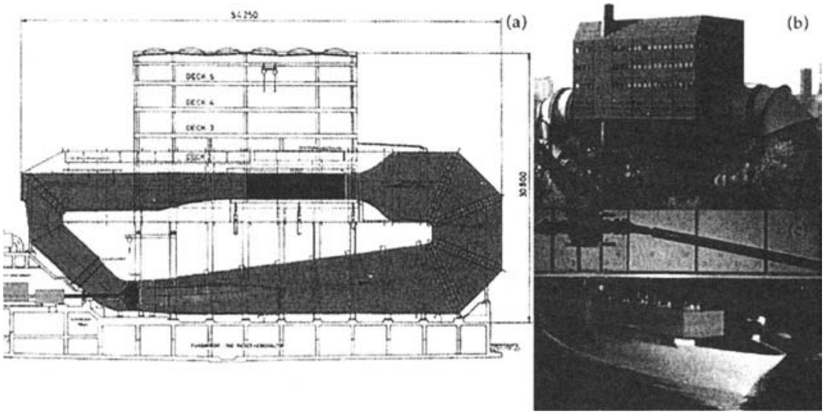



Figure 9.2: The large circulating tunnel of the Versuchsanstalt für Wasserbau und Schiffbau at the Technische Universität in Berlin is the biggest facility of this kind worldwide (a, b). It allows to run simulations with ship-models up to 10 meters length. (c) shows a simulation-run with a model of one of the first super-tankers, the Greek "Tina Onassis". (Source: http://www.tu-berlin.de/vws/)

ilar systems which enable us to perform reliable simulations: (a) by means of reproduction — to which I will refer as *conservative approach*, since as many features of the original system as possible are tried to be 'conserved'; (b) by means of deliberate design and construction — which I will call *constructive approach.*

The conservative approach simply consists in rebuilding the original system as accurately as possible using the same materials the original system is made of. If necessary, the copied system might be scaled by a certain factor. This procedure ensures high structural similarity. Typically, the properties of the simulating system are investigated by purposeful manipulation and intervention wherefore conservative simulations could also be referred to as (indirect) experiments. Figures 9.1 and 9.2 illustrate the conservative approach to simulation by an example from shipbuilding.

The second way to obtain a structurally similar system, namely by the constructive approach, is quite different from the first. Here, the simulating system need not necessarily be built of the same material as the original one.[2] Structural similarity is ensured by the construction of the simulating system according to a plan: The whole simulating system is deliberatively designed in order to realize the same structure as the simulated one and hence to satisfy the prerequisite of structural equivalence. This, of course, requires explicit knowledge ex ante of the structure of the system which is meant to be simulated! Figures 9.3 and

2 In fact, the most comfortable device to 'build' the simulating system under the constructive approach is a computer because of the ease with which it realizes an arbitrary structure.

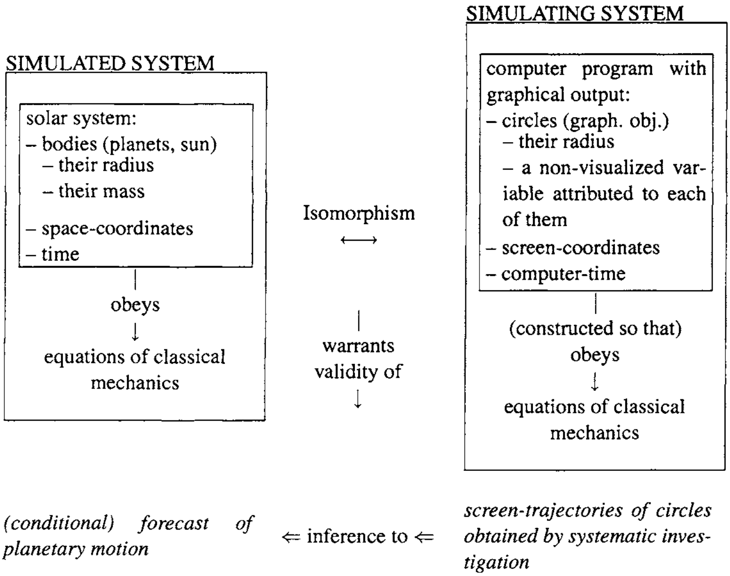



Figure 9.3: The planetary motion in a solar system can be simulated by a computer program, following the constructive approach. In order to make sure that the simulation is reliable, the program has to be designed so as to realize the same formal structure as the simulated system. In particular, the motion of the graphical objects has to be governed by the (translated) equations of classical mechanics. The resulting isomorphism between the two systems warrants the validity of the crucial inference of simulation.

9.4 provide an illustration and an example of this second approach to simulation, respectively.

Before investigating whether these approaches can be applied in economics and climatology, we will reconsider our initial example of nuclear weapon simulation for a moment. Although it involves computer simulation and therefore seems to implement the constructive approach, "experimental capabilities would be crucial", according to a Livermore physicist. "We'd need laboratories where scientists could *scale* nonnuclear experiments *to closely match* weapon physics conditions so they could examine properties at the microstructural level." (cited in Middleton, 2002, my emphasis) Here, the conservative and constructive approach go hand in hand. Over and above that, the computer simulations, as explained on the LLNL-website introducing the Advanced Simulation and Computing Program (ASCI), integrate "data from past nuclear tests with past and present nonnuclear tests, fundamental science and component-level experiments, surveillance of actual weapons withdrawn from the stockpile, and advanced simulations" (Middleton, 2002). Thus it becomes clear that the successful computer simulation of nuclear tests is based on a considerable amount of empirical knowledge includ-

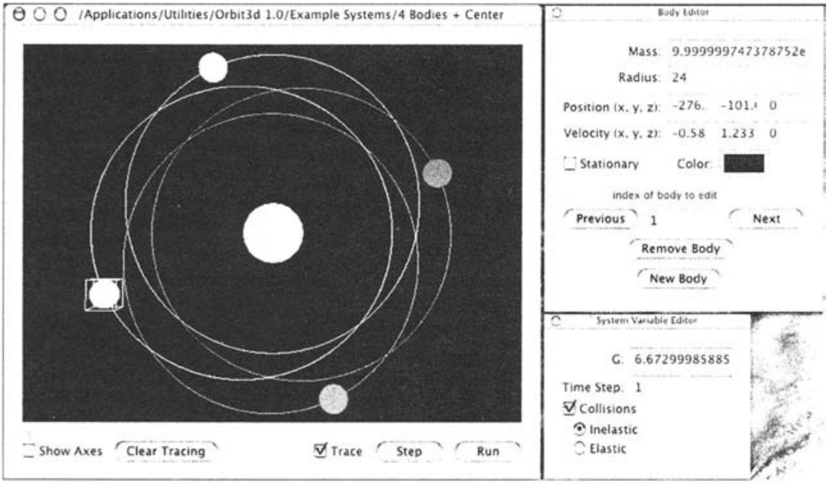



Figure 9.4: This is a screenshot showing the Orbit3D computer program that allows to simulate planetary motion. It is an example for how the right-hand side of figure 9.3 might actually look like.

ing knowledge about detonations of nuclear bombs which, in turn, was acquired by nuclear tests. The reliable simulation of nuclear detonations is only possible because nuclear tests were carried out before. Under this closer inspection, the initial promise of compensating the lack of experiments in the social sciences by simulation thus seems to vanish.

Now, in order to prove that simulation cannot replace experimentation in economics, and to explain macroeconomic forecast failure, I will show in the following for each of the two approaches that (i) it cannot be applied to economics or climatology, and (ii) could it be applied, it would be impossible to obtain interesting insights even into the simulating system. The next section deals with the constructive approach, before the conservative approach is discussed later on.

## 9.3 The constructive approach: computer-simulation in economics and climatology

In this section, I give reasons for two claims: First, it is almost impossible to deliberatively construct systems that are structurally equivalent to either our economic or our climate system. Secondly, if this were possible, we would not be able to obtain any forecasting-relevant insights into these simulating systems which could be applied to the simulated system.

As we have seen above, deliberate construction of reliable simulating systems requires ex ante knowledge of the simulated system's structure. But it is exactly this kind of knowledge we lack in economics and climatology. The vast variety of different economic models (including minimal theoretical, large econometric, as well as multiple-agent models) and climate models (recall the discussion in chapter 4) demonstrates that we do not possess a blueprint of these systems' underlying structures. As long as we lack this knowledge, we are simply not able to deliberatively construct reliable simulations. Furthermore, insights into the original systems can only be obtained by observation of and experimentation with this system. So even if we will have the required knowledge to construct reliable simulations some day, this will be due to experimentation — as is the case for the nuclear weapons example. Simulations are thus far from replacing experiments entirely.

A second argument in favor of the first claim pinpoints technological difficulties one faces when constructing reliable simulations. Just remember that a computer-simulation that integrates all known aspects of our climate would not even run on our fastest super-computers. Maybe the climate as well as the world economy are simply too large to be simulated — maybe a simulating system is required to be of the same size as the climate or the world economy in order to be reliable.

There is a further, very prominent and more fundamental argument that especially applies to economics. It denies that there is something like a stable, time-invariant structure of the economy at all because economic agents' behavior and decision making changes as their expectations change. Such a change of expectations might for instance occur as a result of policy change. Its manifestation is the break-down of statistically estimated economic relationships. This is of course the well-known Lucas-critique reconstructed in a very general form (see Lucas, 1976). As it assumes that the structure of the economy is stated in relationships between aggregated variables of the behavior of individual agents, the challenge the Lucas-critique gives rise to consists in identifying an expectation-change invariant structure of the economy. In the neoclassical research program, this would imply to completely integrate the expectations-forming process into the decision making model: a difficulty we dealt with in chapter 7. The existence of an expectation-change invariant (quantitative) model would imply that there were expectation-change invariant parameters. And such "deep parameters" (Sheffrin, 1996, p. 97) would represent the natural constants of economics. Economists have not yet discovered any such natural constant, and we might doubt whether they ever will.[3]

[3] Although there is no reason to believe they will, it cannot be ruled out either, since there might be an alternative description of economic phenomena, for instance in terms of institutions only, that allows to identify an invariant structure (see also Cartwright, 1999, p. 157). This is, again, the internalist caveat.

Let us turn to the second claim now and assume that we would have a structurally equivalent simulating system at our disposal. I maintain that even this would not be enough to perform reliable simulations because we would not even get the required insights into the simulating system.

The first argument for this second claim scores by simply recalling the discussion on sensitive dependence on initial conditions. If a structural similarity holds between simulated and reliable simulating systems, and if both simulated systems, namely the economy and the climate, exhibit SDIC, and if SDIC is a structural property of a system, then the simulating systems will exhibit SDIC, too. Since the antecedent-conditions are all true, so is the consequence. Thus, if it is impossible to state correct forecasts regarding the original system because of SDIC, it is also impossible to state correct forecasts regarding the simulating system.

A further reasoning adds to the SDIC-argument. Simulations obviously do not solve the data-problem discussed in chapter 6. As we have seen, we ignore the precise state of our economies and our climate. However, any initialization of the simulating system aiming at closest correspondence with the simulated system will be imperfect and the inference from the behavior of the simulating system to the (future) behavior of the simulated one is impeded without accurate data describing the actual state of the simulated system.

## 9.4 The conservative approach: experimental economics

This section deals with the conservative approach and its application to economics. Again, I defend two claims similar to those of the previous section: First, it is impossible to replicate the economy in order to obtain systems suited for simulation. This entails in particular that current laboratory experiments are not (sufficiently) structurally similar to the economic system. Secondly, even if we had successfully copied and scaled the economy, we would not obtain any insights that would improve macroeconomic forecast performance.

Edward Hastings Chamberlin has not only made important theoretical contributions to microeconomics by introducing a model of monopolistic competition, he is also one of the pioneers of experimental economics. In one of the first published articles reporting a laboratory experiment, he states that

> the real world of human beings, firms, markets, and governments cannot be reproduced artificially and controlled. (Chamberlin, 1948, p. 95)

And this coincides with my first claim. Can that be denied at all, or is not it a commonplace, as Chamberlin notes? It probably is. Yet, this statement stands at the beginning and not at the end of the history of experimental economics! How can that be explained? The answer is that laboratory experiments are of interest to microeconomic questions only. What they are meant to simulate are not whole economies but groups of economic agents or single markets. This free standing statement shall now be supported, and discussed in some detail.

Let us see first of all how economic experiments are conducted in general, before we consider an example that successfully applies the conservative approach to a microeconomic problem. The first step in any economic experiment is the recruitment of participants. Conducting experiments with human beings is just the implementation of the general idea that the simulating system should be made of the same stuff as the simulated one. Experimenters seem to prefer students — probably because of their abundance at universities. Vernon Smith, for example, who won the nobel prize in 2002 "for having established laboratory experiments as a tool in empirical economic analysis, especially in the study of alternative market mechanisms" participated as a student in the first experiments conducted by Chamberlin (Holt, 1995, p. 350). The second and most important step is to instruct the participants, i.e. to tell them their action-options and the rewards allocated in function of the individual's and group's behavior. This resembles very much explaining the rules of a game. According to Holt (1995), the rules of Chamberlin's first game were these,

> [Hand] out cards with value and cost information to students [...]. For example, a student receiving a seller card with a cost of \$1 would have the capacity to sell one unit [of an artificial commodity], and the profit on this unit would be the difference between the sale price and the cost of \$1. This seller would have an inelastic supply function with a step at \$1. Similarly, a subject receiving a buyer card with a value of \$2 would have the ability to purchase a unit, and the profit on the unit would be the difference between the value of \$2 and the price paid for the unit. This buyer would have a perfectly inelastic demand for one unit of the commodity at any price below \$2. If this seller and buyer were to arrange a contract for a price of \$1.50, each would earn 50 cents from the trade. Other buyers and sellers can have different values and costs. The market demand and supply function result from a horizontal summation of the individual buyers' and sellers' demand and supply functions. Students were allowed to circulate around the room and arrange trade in a decentralized manner. (p. 350)

In this example, the experimenter completely controls the supply and demand structure of the laboratory market by setting the reservation prices, i.e. printing the cards (see figure 9.5). After instruction of the participants, they can, in a third

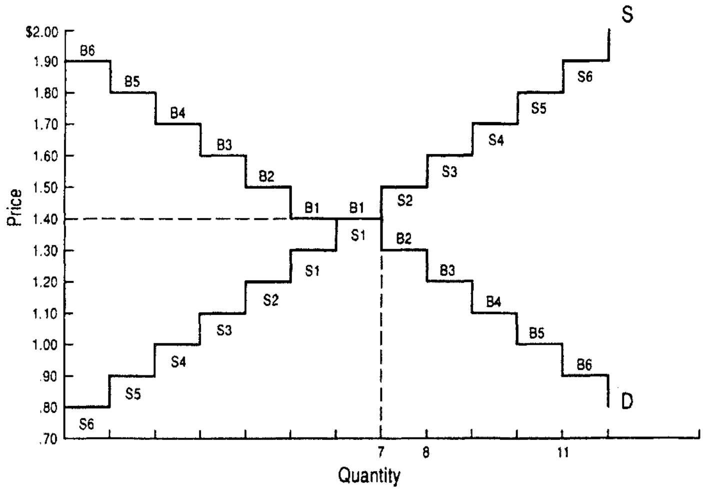



Figure 9.5: These supply- and demand-curves are induced by an experiment involving six sellers (S1 ... S6) and an equal number of buyers (B1 ... B6), each facing different reservation prices and being allowed to make two transactions. (S4, for instance, may sell two units at costs of $1 and $1.70 respectively.) (Source: Holt, 1995, p. 359)

step, engage in the experiment, here: trading, while their behavior is carefully recorded. Nowadays, experiments, in contrast to the one of Chamberlin, use to be carried out with real money in view of ensuring that the instructions are effective and the rewards really determine the participants' incentives.

After these introductory remarks, we are now in a position to review an excellent example of a simulation implementing the conservative approach to simulation in economics. Conducted by Hong & Plott (1982), the economic experiment

> arose in a matter to the Interstate Commerce Commission [... It] had to do with complex posted price markets, and [...] an attempt was made to mirror as closely as possible in the laboratory the industrial structure of the market in question. (Roth, 1995, p. 55)

The simulation served the purpose to decide whether or not barge operators should make their prices public, i.e. post price changes in advance to the Interstate Commerce Commission.[4] In this example, the laboratory market corresponds to what we have called the simulating system and the target market is the simulated one.

> Hong and Plott proceeded to design their experiment around a laboratory market scaled to resemble, in the features mentioned above [e.g.

[4] This Commission had been the oldest regulatory agency in the United States and had primarily been charged with regulating the domestic transport market before it was terminated in 1995.

> the relative size of buyers and sellers, the demand and supply elasticities, the cyclical nature of demand], the market for transporting grain along the upper Mississippi River and Illinois Waterway during the fall of 1970. [...] Aggregate supply and demand functions for the laboratory market were scaled to estimates available for the target market, as was the distribution of large and small firms on each side of the market. The laboratory market was divided into periods representing two weeks of the target market, and the seasonal aspects of the target market were modeled by having demand in the laboratory market scaled to resemble two months of normal demand followed by two months of high demand, followed by two months of normal demand. The experimental design involved running the market under both posted price and negotiated price policies. (Roth, 1995, p. 56)

Given the experimental observation, Hong & Plott (1982) concluded: "The posted price policy causes higher prices, reduced volume, and efficiency losses." (p. 10) It has, subsequently, not been implemented. Although this experiment is relatively complex, it is still a microeconomic experiment since only one single market was investigated. It therefore demonstrates the sheer impossibility to replicate the whole economic system consisting of thousands and thousands of markets like the market for transporting grain along the upper Mississippi River and Illinois Waterway during the fall of 1970.

Before proceeding to a discussion of the second claim, we should briefly consider the relationship between micro- and macroeconomics, and whether improved microeconomic forecast performance entails an improvement of macroeconomic forecast performance. This, however, is not the case since the only macroeconomic theories that are directly related to such microeconomic experiments are those with micro-foundations. Neoclassical macroeconomic models which for instance implement the rational expectations hypothesis might be submitted to experimental tests. And as we have seen in chapter 7 as well as in part 1, these models are of no great importance for macroeconomic forecasting. Still, Roth (1995) makes a more general point by proposing an argument from analogy,

> It may be helpful to point out, by way of loose analogy, that experiments play a role in most of the things biologists do, and, like economists, biologists have a lot of ground to cover, from molecular biology to evolution, to medicine. Experiments can obviously play a very direct role in testing and refining theories of molecular biology, since the phenomena in question can be brought entirely into the laboratory. But although experiments cannot be conducted on the fossil record, evolutionary biologists nevertheless obtain much of their understanding of selection

> and evolution from experiments in microbiology, genetics, and plant breeding. [...] In the same way, economic experiments may play a role not only in testing and refining theories concerned with individuals or small groups, but also concerning questions about large markets, industrial organization, and macroeconomics. (pp. 22f.)

To be fair, we should note that Roth's explicit conclusion does not entail that the predictive performance of macroeconomics will be improved thanks to experimental research. Such a conclusion would indeed not be warranted by this argument as a negative analogy demonstrates: Ever more detailed microbiological knowledge does not and will not necessarily improve the forecastability of an ecosystem or a person's curriculum vitae. Or, to draw another analogy to a science we are more familiar with by now, ever more detailed knowledge of the thermodynamic properties of steam, ozone, carbon dioxide, and so on does not and will not necessarily improve the predictive performance of climatology. The underlying philosophical reason is: Macro-models are not built from bottom up. Reductionism is an inadequate theory of science and it is particularly inadequate when applied rigorously in the context of discovery. Our case studies in climatology have shown that the problem of climatology is not the lack of accurate knowledge on the micro-level. On the contrary, given that some climate models have deliberately violated fundamental physical relationships as for instance the principle of mass- and energy conservation, one might even be tempted to say that empirical adequacy on the one hand and compliance with fundamental laws on the other pull in opposite directions.

As a matter of course, it does not follow from all this that experimental economics could not be a fruitful research program. Like in the natural sciences, economic experimental settings might one day be considered as an interesting research object in their own and thus become the genuine object of inquiry. Experiments, then, would not be required to be structurally similar to any real system. Economists would construct microeconomic nomological machines possibly giving rise to interesting, replicable (and therefore predictable) behavior.

The second claim I maintain in this section needs to be defended yet. It states that, if it were possible to obtain a system suited for simulation by the conservative approach, an investigation into such a system would not bring about forecast-relevant insights. First of all, the problems mentioned in the preceding section, namely the SDIC-argument and the data-problem, apply here, too. However, besides the problems of the constructive approach to economic simulation, the experimental method in economics in general and the conservative approach to simulation in particular face a further obstacle, namely the lack of sufficient control.

Experimenters acknowledge that the participants of economic experiments are sometimes motivated by considerations that are very difficult to control, even in spite of high monetary incentives.[5] Additionally, different participants are motivated by differing considerations which poses an obstacle to the replicability of economic experiments. There is a further aspect of this problem: The background theories underlying the experiments usually presume utility maximizing behavior of economic agents. Therefore, "experimenters have to be alert to deviations from utility maximizing behavior on the part of experimental subjects" (Roth, 1995, p. 79). But such deviations are the rule rather than the exception — a fact supported by many experiments, and which we will seize in the next chapter. Gode & Sunder (1993) pinpoint the problem, ...

> It is not possible to control the trading behavior of individuals. Human traders differ in their expectations, attitudes towards risk, preferences for money versus enjoyment of trading as game, and many other respects. The problem of separating the joint effects of these variations, unobservable to the researcher, can be mitigated by studying market outcomes with participants who follow specified rules of behavior. (p. 120)

... and immediately propose the following solution:

> We therefore replaced human traders by computer programs.

This, of course, means to say goodbye to the conservative and to return to the constructive approach to simulation.

In this section, we have primarily focused on economics neglecting climate experiments. But obviously, the conservative approach cannot be applied successfully in climatology, either. Copying and down-scaling the earth is impossible not only because we lack the necessary technological skills, it is even physically impossible. Even if we were able to assemble a micro-earth, natural laws would hamper a structural similarity to the real earth. A physical model of our planet whose radius is $10^{-3}$ the Earth's radius would give rise to a gravitational field whose strength would equal $10^{-9}$ the Earth's one, it could thus not keep hold of an atmosphere at all.

The overall argument of this chapter can now be stated as follows,

(1) A simulation is reliable if and only if the transfer of insights from the simulating system to the simulated one is truth-preserving. (Definition of reliable simulation)

[5] See for instance Roth (1995, p. 11, pp. 66f.) or Ledyard (1995, p. 115).

(2) The transfer of insights from the simulating system to the simulated one is only truth-preserving if simulating and simulated system are structurally equivalent. (Validity-condition of analogies)

(3) Structurally equivalent systems can be obtained either by the conservative or by the constructive approach to simulation.

(4) The economy cannot be copied and scaled. (Failure of the conservative approach)

(5) In order to design and construct a system $S^c$ that is structurally equivalent to a system $S$, the structure of $S$ has to be known. (Condition of the constructive approach)

(6) The structure of the economy is not known.

(7) If the availability of an object $x$ is necessary for $p$ and $x$ cannot be obtained, then $p$ is false.

(8) *Thus:* Economic simulations are not reliable.

(9) Only reliable simulations give rise to new insights into the simulated system.

(10) Carrying out a simulation does not bring about replicable regularities in the simulated system.

(11) *Thus:* Simulations in economics do neither produce replicable economic regularities nor do they give rise to new insights into the economic system.

(12) The functional role of experiments with regard to the generation of forecasts is to produce replicable regularities and to give rise to new insights into the research object.

(13) A research method $R$ can only be replaced by another research method $R'$ with respect to some goal $G$ if $R'$ can fulfill the same functions as $R$ with regard to $G$.

(14) *Thus:* Simulations in economics cannot replace experiments with regard to the generation of forecasts.

(15) If simulations in economics cannot replace experiments with regard to the generation of forecasts then they do not and will not improve the predictive performance of economics by replacing experiments.

(16) *Thus:* Simulations do not and will not improve the predictive performance of economics by replacing experiments.

## 9.5 A further purpose of simulation: density forecasting

In this last section, a further purpose of simulation shall be discussed: the generation of density and thus probability forecasts. My aim here is to reveal the absurdity of a specific method of establishing probability forecasts in climatology that is based on simulation. It would be a crucial fact for decision making if (macroeconomic or climate) probability forecasts were credible, as we will see later in part 3. Since a credible probabilistic prediction involves objective rather than subjective probabilities (which will also become clear in part 3), we have to sketch an account of objective probabilities.

In one sentence, an objective interpretation of probabilities takes probability statements to be scientifically justified like statements about a body's mass or its charge.[6] As a consequence, how such statements involving objective probabilities can be established and how they have to be understood depends on one's underlying general philosophy of science, in particular one's methodology and account of how theoretical terms relate to observation. Accordingly, there is no single philosophical interpretation of objective probabilities, but a large variety.

Van Mises frequency theory of probabilities is based on operationalism. Whereas his theory has been widely influential, philosophers of science largely reject operationalism and accordingly develop alternative philosophical theories of probability.[7] Gillies (2000, p. 114) proposes to refer to all such alternatives as propensity theories. A particular propensity theory is Nancy Cartwright's view which is based on her theory of science and developed from Ian Hackings analysis,

> Ian Hacking in *The Logic of Statistical Inference* taught that probabilities are characterized relative to chance set-ups and do not make sense without them. [...] A chance set-up is a nomological machine for probabilistic laws, and our description of it is a model that works in the same way as a model for deterministic laws. (Cartwright, 1999, p. 152)

According to this account, objective probability statements characterize well specified, shielded and controlled set-ups only. Only nomological machines which are adequately described by a probabilistic model can be characterized by objective probabilities.

Yet, this idea that objective probabilities "are characterized relative to chance set-ups and do not make sense without them" is central to most propensity theories (Gillies, 2000, p. 115). The following reasoning relies on it.

6 Hence, an objective interpretation of probabilities does not automatically imply a probability realism, i.e. the metaphysical claim that the world as such is non-deterministic and governed by probabilistic laws, for that depends on whether scientific realism is adopted or not.

7 See Gillies (2000, pp. 125).

The propensity account of probabilities has important implications for density forecasting by means of simulation. As probabilities are bound to and determined by a system's underlying structure and since reliable simulation is characterized by the structural similarity of simulating and simulated system, the inference from probabilities characterizing the simulating system to those of the simulated one is valid. In other words: Density functions of systems can be estimated by reliable simulation and *reliable* simulation only.[8]

However, there are no reliable simulations of our climate system or our economy as I have argued throughout this chapter. The road that leads to density forecasting via simulation thus seems to be blocked. Nevertheless, some climate scientist do consider probability forecasting by simulation as feasible. Here is a quote from the IPCC's Third Assessment Report,

> In climate research and modeling we should recognize that [...] the long-term prediction of future climate states is not possible. The most we can expect to achieve is the prediction of the probability distribution of the system's future possible states by the generation of ensembles of model solutions. (IPCC, 2001*b*, p. 774)

This methodology of "ensemble predictions" is specified by another author of the TAR,

> [Uncertainties] can be partially quantified from ensembles of climate change integrations, made using different models starting from different initial conditions. They necessarily give rise to probabilistic estimates of climate change. (IPCC, 2001*b*, p. 419)

More specifically, what these climatologists propose amounts to the claim that the simulation results presented in figure 4.14 induce a density forecast for the sea level rise in the next hundred years. Accordingly, this figure may be interpreted as stating a sea-level rise of 0.4 m to be the most probable outcome while sea level rises of 0.8 m or 0.1 m are also possible but much less probable.

Still, it should be mentioned that the TAR also expresses doubts relating to this methodology. The IPCC (2001*b*) critically notes,

> An important question is whether a multi-model ensemble made by pooling the world climate community's stock of global models adequately spans the uncertainty in our ability to represent faithfully the evolution of climate. (p. 423)

8 By the way, a purely subjective interpretation of probabilities could not warrant this inference.

And the IPCC (2001*b*) skeptically remarks in its conclusion

> that it is important to assign probabilities to projections, but this requires a more critical and quantitative assessment of model uncertainties than is possible at present. (p. 681)

The following is an attempt to make the assumptions underlying the ensemble-methodology explicit and to strengthen the critiques. Let $C$ denote a certain climate event, for instance the sea-level rise of 0.6±0.01 m during the next hundred years. Assume, *for the sake of argument*, that for each simulation $s_i$ of our ensemble $\{s_1, \dots, s_n\}$ the probability of $C$ is estimated, i.e. we know the conditional probability of $C$ given that $s_i$ is a reliable simulation, $\mathrm{P}(C|S_i)$.[9] Now, the proponent of the ensemble-methodology claims that this allows us to estimate the unconditional probability of $C$, namely $\mathrm{P}(C)$. The calculation implicitly performed seems to be,

$$\begin{aligned}\mathrm{P}(C) &= \mathrm{P}(C) \cdot \underbrace{(\mathrm{P}(S_1) + \ldots + \mathrm{P}(S_n))}_{=1} \\ &= \mathrm{P}(C\&S_1) + \ldots + \mathrm{P}(C\&S_n) \\ &= (\mathrm{P}(S_1) \cdot \mathrm{P}(C|S_1)) + \ldots + (\mathrm{P}(S_n) \cdot \mathrm{P}(C|S_n)),\end{aligned}$$

which makes use of the definition of conditional probabilities,

$$\mathrm{P}(A|B) := \frac{\mathrm{P}(A\&B)}{\mathrm{P}(B)}.$$

Now, this calculation assumes that (i) $\sum \mathrm{P}(S_i) = 1$ and (ii) $C$ and $S_i$ are independent for all $i$ ($\mathrm{P}(C)\mathrm{P}(S_i) = 1$). The first assumption has been questioned in the above quotes. But now consider that both the conditional probabilities $\mathrm{P}(C|S_i)$ and the probabilities $\mathrm{P}(S_i)$ have to be known in order to calculate the unconditional probabilities $\mathrm{P}(C)$. That is exactly where my critique relying on Cartwright's analysis of objective probabilities (or more generally propensity theories) attacks. I maintain that there is nothing like $\mathrm{P}(S_i)$, i.e. the objective probability of $s_i$ being a reliable simulation! How could we estimate the probability that a simulation is structurally similar to a real system? — Similarly: What is the probability that a certain model, for instance the Kaldor model, is the true model of the system under investigation? — Where is the nomological machine which gives rise to these probabilities?! These 'probabilities' are at most subjective degrees of belief.[10] A methodology which assumes that the probabilities of

9 In doing so, we make the strong assumption that the uncertainty relating to initial and boundary conditions is objectively quantified.

10 "At most" because subjective probabilities are defined as consistent betting quotients and

simulations being structurally similar to the real system can objectively be estimated seems just crazy to me. Thus, let me repeat: What we need for probability estimation is one reliable simulation and pooling different simulations will never allow for estimating a single density function. This holds for climatology as for any other science.

it is not unanimously agreed on what type of beliefs one can meaningfully bet at all. As Stegmüller (1973, p. 232) points out, de Finetti who initially developed the subjective interpretation of probabilities insisted that if a hypothesis cannot be verified by observation "its probability is meaningless." It follows that according to de Finetti one cannot attach subjective probabilities to the alternative climate models.

# Chapter 10

# Unrealistic-assumption explanations


## Summary

Macroeconomic and climate forecasts are largely based on unrealistic assumptions. This fact as such, however, does not explain forecast failure as the counterexamples of Larry Laudan's pessimistic meta-induction show. Once this is established, this chapter develops a more sophisticated explanation of forecast failure involving unrealistic assumptions by a critical discussion of the molecular kinetic deduction of the ideal gas law and its improvement. But, finally, this type of explanation cannot be applied in economics.


## 10.1 Unrealistic assumptions of macroeconomic forecasts

I suggested in chapter 7 that the basic assumptions of economics as for instance the expected utility maximization or the rational expectations hypothesis (EUH and REH) are too abstract to have any empirical content and that they should be interpreted as general principles stating how to construct particular models. This said, we all make the every-day experience that standard models built according to EUH and REH are constantly violated. We and our fellow citizens simply do not take our decisions in order to (or so as if we) maximize expected utility when utility is defined as a monotone function of consumption with a positive, declining slope. Likewise, our expectations are not rational in the sense that we consider any available information when forming them.

The common sense knowledge that the basic assumptions of mainstream economics are literally false, or, as people use to say, unrealistic, is confirmed by empirical research. A large part of economic experiments is devoted to testing these assumptions. Having presented the growing evidence that was accumulated during the last decades, Roth (1995) summarizes,

> As the number of replicable violations of utility theory, and even of more basic models of rational choice, has grown, a question we frequently hear from some of our psychologist colleagues, and one that we can reasonably ask ourselves, is "what accounts for economists' reluctance to abandon the rational model, despite *considerable contradictory evidence*?" (p. 76, my emphasis)

Or, as Roth's section-title puts it, "Why haven't these demonstrated anomalies swept away utility theory?".[1] The REH has to face contradictory — while less conclusive — empirical evidence, too. Concluding on overlapping generations experiments, Ochs (1995) writes that one might

> read this evidence as supporting the methodological position that one cannot avoid the problem of modeling the process of expectations forming [by using the REH short-cut] and hope to have an empirically valid model. (p. 209)

Now, in addition to these experimental findings, statistical data may give a hint on whether basic applications of REH are adequate or not. REH has been submitted to statistical testing by checking individual forecasts for basic criteria of rationality as unbiasedness and efficiency. Sheffrin (1996) summarizes these findings as follows,

> On the whole, survey data do not support the rational expectations hypothesis. But the rather mixed evidence from all these diverse surveys does not necessarily imply that the rational expectations hypothesis will be of only limited use for economics. (p. 21)

Since we have not questioned yet whether REH is of use but only whether its basic applications accurately describe an agent's expectations, the quote should be counted as additional evidence against REH's realism.

Paraphrasing Cartwright (1983*b*), one might conclude that the basic laws of economics lie. Or, in the words of Paul Krugman, "all our models involve silly assumptions" (Krugman, 1998, p. 150). In so far as macroeconomic forecasts make use of economic theory, they are based on unrealistic assumptions where "unrealistic" means false according to the accepted theory, empirical evidence or common sense. However, as we have learned in part 1, the majority of macroeconomic forecasts do not make direct use of economic theory, i.e. are non-econometric. But the assumptions of alternative methods are no more realistic than those of economic theory. Univariate ARIMA forecasts, for instance, assume that future values of the

[1] This question will in fact be answered 'by the way' in due course of this chapter.

predicted variable depend on the variable's values of the last few years only, denying any other influences, whereas naïve methods make even stronger assumptions. Indicator forecasts assume suspicious lead- and lag-relationships to hold and judgmental procedures, finally, presuppose that the forecaster is a reliable clairvoyant. The question of this chapter is: Does this obvious unrealism explain forecast failure?

## 10.2 Reconstructing the unrealistic-assumption explanation

Before attempting to reconstruct the unrealistic-assumption explanation we must briefly clarify what it means to explain something at all. This is not the first time that we deal with explanation in our inquiry as all the preceding chapters of part 2 were meant to explain forecast failure. What did we do in these chapters? Actually, we have discussed arguments in favor of the conclusion that macroeconomic forecasts are not successful (or in favor of any other more specific observation of part 1 which we have tried to explain). In so far as these arguments were plausible, we have considered them as good explanations. Plausible arguments, according to our practice, are explanations. Now, I also want to defend the inverse claim that any explanation is a plausible argument. First of all, to explain a certain phenomenon always means to explain why a statement (that asserts the occurrence of the phenomenon) is true. We do not explain phenomena per se but statements describing phenomena. This is illustrated by the fact that explaining is essentially answering a why-question and "Why ..." is followed by a sentence. Now, how do we explain a certain statement? How do we answer why-questions? Since explaining is a cognitive enterprise, this is necessarily done by uttering further statements. And their truth is a necessary condition for accepting the explanation, or the answer, respectively. Yet giving true statements is clearly not a sufficient condition for answering a why-question satisfactorily. Additionally, the explaining statements have to stand in a certain relation to the explained one. However, the only epistemologically relevant relations between statements are the logical relations. And since explaining does not involve establishing contradictions or mere coherence, the relation to be established is an inferential one. Consequently, an explanation is an argument. Taking all this together, an explanation of a statement $q$ is given by a plausible deductive or inductive argument (for instance an inference by analogy) with conclusion $q$. As far as I can see, the inferential conception of explanation is not an alternative to a causal or a nomological or any other account of explanation, but it is the more fundamental conception that underlies all these alternative accounts of explanation — which are still subject of philosophical debate. Although some philosophers strictly oppose the inferential conception of explanation,[2] there is no room here to discuss this in more detail. I take the inferential conception for granted.

This said, the unrealistic-assumption explanation of forecast failure can be reconstructed as the following argument with $F$ denoting a forecast and $A$ an assumption,

(1) $F$ assumes $A$.
(2) $A$ is unrealistic.
(C) *Thus:* $F$ is not successful.

Clearly, this argument is neither a valid deductive nor an inductive argument and the so-reconstructed explanation fails. But there is an obvious (trivial) way to fix it by introducing a further premiss[3]:

(3) If $F$ assumes $A$ and $A$ is unrealistic then $F$ is not successful.

Now, (1), (2) and (3) entail (C), indeed, and the argument is valid. But is it sound? Premisses (1) and (2) are true according to the problem's definition. And number (3)?

## 10.3 Criticizing the unrealistic-assumption explanation

This section's argumentation aims at showing that (3) is not true and the above reconstruction of the unrealistic-assumption explanation fails. The main work has already been done by somebody else whose results will be presented here. This person is Larry Laudan and the work we refer to is his well-known pessimistic meta-induction (see Laudan, 1981)[4].

The background of Laudan's work is quite different from ours: In the last decades, one of the predominant debates in philosophy of science has been the

[2] See for instance Salmon (1989, p. 174).

[3] Adding this premiss does not force us to subscribe to a certain conception of explanation more specific than the inferential one. Lycan (2002, p. 411) lists five different paradigms of explanation: subsumption, showing-to-be-expected, causal, filling-the-gap-in-understanding and reduction where the last may be omitted in our context. By altering the comment that we give to (3), our explanation can be attributed to any of the remaining four paradigms. If we consider (3) as a general regularity, it is an explanation by subsumption. Were (3) an empirically successful regularity, (C) would be shown to be expected. If (3) is an acceptable statement, the explanation would fill a gap in understanding by showing up inferential relationships. And finally, (3) may even be interpreted as stating a causal relationship: If $F$ assumes $A$ and $A$ is unrealistic, then $A$'s falsity causes $F$ to fail. (Under this lecture, an implicit fourth premiss is involved.) In sum, and this underlines the generality of the inferential conception, we may remain neutral with respect to the alternative accounts of explanation.

[4] Page numbers according to the reprint in Papineau (1996).

quarrel about scientific realism, i.e. the thesis that our best scientific theories are at least approximately 'true'[5] and do properly refer. In the 1970s, some philosophers, Hilary Putnam (Putnam, 1975) first and foremost, proposed an important argument in favor of scientific realism: the so-called no-miracle argument. It claimed that scientific realism is the best (because: only) explanation of the success of science and that its success will remain a miracle under any other account of science. Laudan (1981) presents a striking attack of this no-miracle argument. His strategy is to show that scientific realism does not explain the success of science at all. It is not his whole reasoning we are interested in but rather one particular step: the disproof of the claim,

(T2) If a theory is explanatorily successful, then it is approximately true. (p. 118)

A theory's success is, however, not limited to giving satisfactory explanations. More general, a theory is successful "if it makes substantially correct predictions, if it leads to efficacious interventions in the natural order, and if it passes a battery of standard tests" (Laudan, 1981, p. 111). Under this reading, (T2) turns out to be the general version of the contraposition of (3). This raises the hope that Laudan's disproof of (T2) may also establish the falsity of (3) although it does not necessarily do so. From a logical point of view, his argument is as simple as effective: It just consists in many counter-examples. A counter-example to (T2) is a theory that is successful but not approximately 'true'. Given that genuine reference is — at least for the realist — a necessary condition for truth, it suffices to enumerate successful theories that fail to refer.

> Now, what the history of science offers us is a plethora of theories that were both successful and (so far as we can judge) non-referential with respect to many of their central explanatory concepts. (Laudan, 1981, p. 121)

Having discussed the ether-theories of the 19th century as a first example in some detail, Laudan extends his list of counter-examples:

> the crystalline spheres of ancient and medieval astronomy;
>
> the humoral theory of medicine;
>
> the effluvial theory of static electricity;
>
> 'catastrophist' geology, with its commitment to a universal (Noachian) deluge;
>
> the phlogiston theory of chemistry;

[5] In a metaphysical sense.

> the caloric theory of heat;
>
> the vibratory theory of heat;
>
> the vital force theories of physiology;
>
> the electromagnetic ether;
>
> the optical ether;
>
> the theory of circular inertia;
>
> theories of spontaneous generation.
>
> This list, which could be extended *ad nauseam*, involves in every case a theory that was once successful and well confirmed, but which contained central terms that (we now believe) were non referring. [...] But we need not limit our counter-examples to non-referring theories. There were many theories in the past that (so far as we can tell) were both genuinely referring and empirically successful which we are none the less loath to regard as approximately true. (pp. 121f.)

For Laudan's examples are theories (i) that involve — relative to nowadays accepted theories false and thus: — unrealistic assumptions, and (ii) whose success consisted mostly and at least partly in predictive success, they are, as we hoped, not only counter-examples against (T2) but also against (3).[6]

This evidence against (3) may be augmented by examples taken from our own case studies. The ideal gas law, for instance, can be derived from unrealistic assumptions but has been successful for quite a long time as we saw in chapter 4. Or consider climate models! To call certain of their assumptions unrealistic would even be an understatement. Flux adjustments introduce ad hoc assumptions that contradict the most basic principles of physics as energy- and mass-conservation. But maybe climate models are no good counter-example for they are hardly successful. The point, however, is that introducing these unrealistic assumptions *improves* the overall predictive performance. In this respect, climate models represent an even stronger counter-example than Laudan's examples since forecasts based on unrealistic assumptions may not only be successful but even more successful than their realistic counterparts.

Laudan's pessimistic meta-induction plus our own modest record of examples altogether imply that unrealistic assumptions do not explain forecast failure the way suggested in the preceding section.

---

6 Notice that they are also counter-examples against (3) in any of its different readings according to footnote 3.

## 10.4 An alternative explanation involving unrealistic assumptions

The failure of the explanation outlined above does of course not imply that any explanation involving unrealistic assumptions necessarily fails. In fact, what I want to show in the remainder of this chapter is that unrealistic assumptions do have a role to play in explaining forecast failure as well as in improving forecasts whereas both are closely related. While simply stating that a forecast's assumptions are false is not sufficient for explaining forecast failure, a (modified) unrealistic-assumption explanation, in order to be successful, has to show additionally in how far the false assumptions were *responsible* for the forecast's failure. A more detailed case study about a once successful theory involving unrealistic assumptions and its improvement shall pace the way to a better understanding of how such a modified unrealistic-assumption explanation might work.

The case we will have a closer look at is the ideal gas law. It had been considered as empirically quite successful in the first half of the 18th century although some deviations from this law were known by then.[7] In the mid of the 19th century, Rudolf Clausius showed that it can be derived within a molecular-kinetic theory of heat. In order to deduce the ideal gas law, serious simplification, i.e. unrealistic assumptions, have to be made. Namely,

A1) Molecules are Newtonian mass-points without volume, in particular, there are no collisions between molecules.

A2) There are no inter-molecular forces.

When applying Newtonian mechanics to these mass points, the ideal gas law can be derived with some further assumptions as the following reconstruction of a textbook-deduction demonstrates.

Let us set the stage: Consider a cubic box with edge-length $a$ and a coordinate system parallel to the box's edges. The box shall be filled with $N$ homogeneously distributed molecules of a certain gas with molecular mass $m$. These molecules are moving around whereas the velocity-distribution is assumed to be isotropic. This means: For any two directions and any given speed, there is the same number of molecules flying with this speed in both directions.

Our first and principal task is this: to calculate the momentum exerted on the right side wall perpendicular to the $x$-axis by impacts of molecules that hit this wall. In doing so, we only have to consider those molecules that move towards this wall (the right-moving molecules), i.e. those with a positive $x$-component of velocity. As the velocity distribution is isotropic, there are $N/2$ such molecules. Let us name them: The easiest way is to number them by an index $i = 1 \ldots N/2$. Now, $v(i)$ shall denote the absolute velocity of the $i$th molecule and $v_x(i)$, $v_y(i)$, $v_z(i)$ the corresponding components.

---

7 See chapter 4.

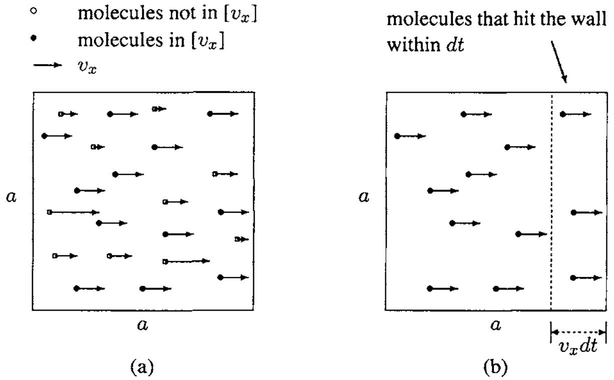



Figure 10.1: Illustrations of the molecular kinetic deduction of the ideal gas law.

As a final preliminary step, we define some structure upon the set of the right-moving molecules. This is achieved by grouping them according to their respective speed: We define an as-fast-as relation between molecules: $i$ as-fast-as $j$ iff $v_x(i) = v_x(j)$. As-fast-as is an equivalence relation inducing a partition on the set of all right-moving molecules with velocity-classes $[v_x]$ including all molecules with (x-component) velocity $v_x$, see figure 10.1a. Moreover, let $|v_x|$ denote the number of right-moving molecules with speed $v_x$.

All this said and done, we can eventually start by calculating the momentum $dp_{v_x}$ imparted by the molecules in the velocity-class $[v_x]$ during a time-interval $dt$ on the right-side wall. Each of these molecules imparts, according to momentum conservation (assuming that collisions are elastic), a momentum of $2mv_x$ when it hits the wall. But how many molecules of velocity-class $[v_x]$ do hit the wall during $dt$? All those that are close enough to reach the wall within $dt$ as illustrated in figure 10.1b. Since the molecules are, by assumption, homogeneously distributed, this is the fraction $v_x \cdot dt/a$ of all molecules in $[v_x]$,

$$\begin{aligned} dp_{v_x} &= \underbrace{\left(\frac{v_x \cdot dt}{a} \cdot |v_x|\right)}_{\#\text{ of hits in } dt} \cdot 2mv_x \\ &= \sum_{i \in [v_x]} \frac{2m \cdot dt}{a} \cdot v_x(i)^2 \end{aligned}$$

The total momentum imparted on the wall in $dt$ is the sum of the momenta imparted by the molecules of each velocity class,

$$\begin{aligned} dp &= \sum_{v_x} dp_{v_x} \\ &= \sum_{v_x} \sum_{i \in [v_x]} \frac{2m \cdot dt}{a} \cdot v_x(i)^2 \\ &= \sum_i \frac{2m \cdot dt}{a} v_x(i)^2 \end{aligned}$$

$$= \frac{2m \cdot dt}{a} \sum_i v_x(i)^2 \tag{10.1}$$

Where $\sum_i$ sums over all $N/2$ right-moving molecules. Next, we have to consider the other velocity-components. Because the velocity distribution is isotropic (all directions are equivalent), the squared velocities sum up to the same number for each component,

$$\sum_i v_x(i)^2 = \sum_i v_y(i)^2 = \sum_i v_z(i)^2.$$

By applying Pythagoras, i.e.

$$\forall i : v(i)^2 = v_x(i)^2 + v_y(i)^2 + v_z(i)^2,$$

we obtain,

$$\begin{aligned} \sum_i v_x(i)^2 &= \frac{1}{3} \sum_i v(i)^2 \\ &= \frac{1}{3}\left(\frac{1}{2} N \overline{v^2}\right), \end{aligned}$$

where $\overline{v^2}$ is the mean square velocity of the molecules. Substituting this into (10.1) yields,

$$\begin{aligned} dp &= \frac{2m \cdot dt}{a} \frac{1}{3}\left(\frac{1}{2} N \overline{v^2}\right) \\ &= \frac{1}{3} \frac{dt}{a} N \cdot m\overline{v^2} \end{aligned} \tag{10.2}$$

As the situation is perfectly symmetric, this is the momentum imparted on any of the six walls of the box. From now on, the derivation is straightforward. Since force is the time-derivative of momentum, (10.2) allows us to calculate the force acting on each wall,

$$\begin{aligned} F &= \frac{dp}{dt} = \frac{1}{dt}\left(\frac{1}{3} \frac{dt}{a} N \cdot m\overline{v^2}\right) \\ &= \frac{1}{3} \frac{1}{a} N \cdot m\overline{v^2}. \end{aligned}$$

The pressure $P$ of the gas is the force divided by the area upon which it acts, namely the wall's area $a \cdot a$, thus

$$P = \frac{F}{a^2} = \frac{1}{3} \frac{1}{a^3} N \cdot m\overline{v^2}.$$

And because $a^3$ is the box's volume $V$ this can be arranged to,

$$PV = \frac{1}{3} N \cdot m\overline{v^2}.$$

Finally, the application of the molecular kinetic theory of heat, namely the identification of (absolute) temperature $T$ with mean kinetic energy of the substance's molecules: Setting

$1/2m\overline{v^2} = 3/2kT$ where $k$ is a constant (the so-called Boltzmann constant), we obtain the ideal gas law,

$$PV = NkT,$$

or, defining the universal gas constant $R$ as the product of $k$ and the number of molecules per mol (Avogadro's number, $N_a$),

$$PV = nRT, \tag{10.3}$$

with $n$ denoting the number of moles.

As mentioned above, deviations from the ideal gas law were already known when Clausius presented his molecular-kinetic deduction. However, empirical evidence against the ideal gas law grew significantly in the second half of the 19th century due to the works of Andrews, Amagat and Cailletet. As a response, van der Waals proposed a modified gas law that was supposed to be compatible with the empirical evidence. But it is a philosophical commonplace that there are infinitely many alternative (gas) laws that are compatible with the necessarily limited amount of data. So how did van der Waals choose among the alternatives? At this point, the unrealistic assumptions of the ideal gas law (A1) and (A2) became extremely helpful. Van der Waals dropped those assumptions, replaced them with new ones, and modified the gas law accordingly. In particular, he assumed that

A1′) Molecules have a specific volume $V_M$.

A2′) There are attracting inter-molecular forces.

(A1') entails that the free volume, i.e. the volume a molecule can occupy, is not equal to the container's volume as molecules may not overlap each other. Since each molecule blocks a space roughly equalling four times its own volume, the free volume is,

$$V_{\text{free}} = V - 4NV_M = V - n\underbrace{4N_aV_M}_{=:b} = V - nb. \tag{10.4}$$

Van der Waals maintained that what the ideal gas law actually talks about is the free volume and the latter must be substituted to $V$.

In addition, (A2') implies that molecules are slowed down shortly before hitting the walls (and thereby inducing the pressure). Also, such hits do occur less frequently than under the absence of any inter-molecular attraction. Consequently, the observed pressure is less than the pressure as determined by the ideal gas law. Van der Waals proposed the following correction term involving a further parameter $a$,

$$P = \frac{nRT}{V_{\text{free}}} - \underbrace{\frac{an^2}{V^2}}_{\text{correction}} . \tag{10.5}$$

Taking (10.4) and (10.5) together yields the van der Waals equation,

$$(P + \frac{an^2}{V^2})(V - nb) = nRT. \tag{10.6}$$

These modifications of the ideal gas law not only led to better predictive performance but also explained the ideal gas law's failure. First of all, for high volumes and low pressures, the ideal gas law is the limiting case of the van der Waals equation. This follows from the fact that (A1) and (A2) are fairly correct approximations for high volumes and low pressures. However, with rising pressure and decreasing volume, inter-molecular forces and the molecules' volumes may no longer be neglected. The modified theory seems to explain the ideal gas law's success as well as its failure. This becomes strikingly clear when we calculate the ideal gas law's forecast error by making use of the van der Waals equation! Let us assume that (10.3) is used to predict the temperature of a gas submitted to ever growing pressure. Consequently, the predicted temperature is,

$$T^F = \frac{PV}{nR},$$

whereas the correct value according to (10.6) equals,

$$T = \frac{1}{nR}(P + \frac{an^2}{V^2})(V - nb).$$

This allows us to determine the forecast error,

$$\begin{aligned} e = T^F - T \quad &= \frac{PV}{nR} - \frac{1}{nR}(P + \frac{an^2}{V^2})(V - nb) \\ &= \frac{1}{R}(bP + \frac{an}{V}(\frac{bn}{V} - 1)). \end{aligned} \tag{10.7}$$

which is growing in $P$ and declining in $V$. Hence, we did not only explain the predictive failure of the ideal gas law in general but also its quantitative forecast errors.

Let us look back at the last main steps of our case study to understand what exactly ensures that this kind of explanation *explains*, and where the difference to the unrealistic-assumption explanation we discussed in the previous sections lies. As a first step, the unrealistic assumptions were refuted and replaced by alternative assumptions which, in turn, induced an alternative model, $\mathcal{M}$. Forecasts based on $\mathcal{M}$, and thus on the modified assumptions, were more successful than the original ones. Finally, and this seems to be the most important thing, $\mathcal{M}$ explained (implied) the failure of the forecasts based on the old assumptions. The general form of the explanation is,

(1) $F$ assumes $A$.
(2) $A$ is unrealistic.
(3') The model $\mathcal{M}$ warrants: If a forecast assumes $A$, then it brings about systematic forecast errors.
(4) $\mathcal{M}$ is (accepted as) true.
(C) THUS: $F$ brings about systematic forecast errors.

Having a closer look at this explanation, we realize that (2) is no longer needed! Indeed, the explanation does not really seem to be an unrealistic-assumption explanation but rather looks like an explanation of forecast failure conditional on a certain model $\mathcal{M}$. However, if we take it for granted that a true model does not bring about systematic forecast errors, then (3') and (4) imply (2), presupposed that $\mathcal{M}$ has to assume either $A$ or not-$A$. Although $A$ being unrealistic is no premiss which is needed for the argument to be valid, it is a necessary condition for its soundness, i.e. for its premisses to be true. In this sense, we have finally identified a type of successful unrealistic-assumption explanation. It is, however, much more sophisticated and demanding than the explanation-type we discussed originally.

Let us finally apply our findings to macroeconomics: The unrealistic assumptions of economics do not per se explain macroeconomic forecast failure. A satisfying unrealistic-assumption explanation analogous to the above type would require a model that modifies these assumptions, that gives rise to (at least relatively) successful predictions and is therefore accepted as true, and that explains why the original assumptions are responsible for macroeconomic forecast failure. But such an empirically adequate alternative to classical economics is yet to emerge. However, this still leaves room for other successful explanations of macroeconomic forecast failure which involve unrealistic assumptions. And to the extent that the explanations discussed in the previous chapters refer to problematic assumptions underlying some macroeconomic forecasts as for instance the low complexity of predictive models or the imprecision of the predictive data, even they might be considered as unrealistic-assumption explanations in the widest sense, too.

# Part III

# Living with the predictive limits of economics

# Chapter 11

# Consequences for traditional decision making


## Summary

Traditional accounts of rational decision making cannot be applied in the field of macroeconomic policy-making because macroeconomic forecasts fail consistently. Having elaborated a minimalist notion of instrumental rationality, this chapter shows in particular that neither the deterministic nor the probabilistic account of rational choice can be applied. Furthermore, rational decision making cannot be based on causal knowledge, either. Eventually, macroeconomic policy-making is characterized as decision making under uncertainty and ignorance.


## 11.1 Instrumental rationality

The concept of (instrumental) rationality is at the heart of our private and social lives. It has a guiding function, and is structuring our individual as well as on collective behavior by providing a procedure for deliberate choice between alternative actions[1]. In its most general version, this procedure consists in assessing the logical connections between alternative choices on the one side and the agent's background knowledge[2] on the other. A choice is rational if and only if it is consistent with the background knowledge. This abstract definition represents a logico-semantical account of rationality since it only takes the logical relations between a strategy's description and some assumed background knowledge into account.

In so far as we can ascribe some shared background knowledge to a group of agents, this account does not only apply to individual but also to collective

[1] Or 'strategies', as I will say below.
[2] Including descriptive as well as normative beliefs.

decision making. It is thus not merely an analysis of individual rationality. We do for example apply the above definition of rationality to group decisions when evaluating the policy of 'the' US-administration ex post ("Why do they oppose the ICJ?" — "They probably belief that ... "), or deliberate what a government should do. Still, we have to assume a homogeneous body of background knowledge and our account of rationality does not provide a procedure for collective rational choice, if multiple agents disagree with respect to their background knowledge. In the light of this constraint, we give an account of "atomistic rationality" (Jaeger et al., 1998).

Our logico-semantical approach implies that it depends on a decision's description, its interpretation whether it is rational or not; for not the strategy as such but only its description is logically connected with other statements. As long as the problem is sufficiently simple, descriptions are definite; for complex strategies, however, several distinct descriptions might be appropriate — an observation which will be taken up in chapter 13. For the time being, we shall assume that the strategies' descriptions are given unambiguously.

If a strategy is rational (to adopt) as long as it is consistent with an agent's background knowledge, the idea of rationality can fulfill its guiding function only if the strategies on the one hand and the agent's background beliefs on the other hand are not logically independent. Otherwise, the decision for *any* strategy would be compatible with her beliefs, which would not be very helpful for making a choice. The background knowledge must therefore include statements which connect strategies and beliefs. I will call such statements *bridge-principles*. Bridge-principles differ in connecting different *types* of beliefs to strategies, and hence make specific requirements regarding the background knowledge. Unless these *epistemic requirements* are fulfilled the alternative strategies will be logically independent from the agent's beliefs in spite of the bridge-principle, and again the idea of rationality cannot fulfill its normative function. The sections below will classify and analyze different procedures of rational decision making according to the bridge-principles involved. I will argue that, given the limits of economic forecasting, the epistemic requirements of standard patterns of rational conduct are unsatisfiable.

Before doing so, let us see how the above definition of rationality relates to, agrees with and differs from other explications of this important idea. Suppose that we had sufficiently strong and satisfiable bridge-principles so that in every relevant decision making situation only one alternative strategy is consistent with the background beliefs and thus rational. Under such circumstances, a decision for a certain strategy is rational if and only if the strategy can be deduced from the agent's beliefs, including the bridge-principles, i.e. if there is a warrant for it. Now, what if one redefines rationality accordingly? Clearly such a warrant-definition of rationality is stronger than our above definition. In particular, it implies that there is (at most) one rational choice for any decision problem. Yet that seems too strong an implication for there are situations such as chance games where alternative, mutually inconsistent strategies are, according to our intuitions, rational — at least not irrational. Facing such counter-examples, the warrant-definition might be saved by claiming that the categories of 'rational' and 'irrational' decisions are not collectively exhaustive: Accordingly, a decision is rational if it is warranted by the agent's beliefs, it is irrational if it is inconsistent with the agent's beliefs, and it is neither rational nor irrational if it is consistent with but not deducible from the background knowledge.

The warrant-definition comes quite close to Habermas' explication of instrumental rationality. However, Habermas does not only require the action to be warranted but also the belief on which it is founded, for instance the assumption that the action is efficient,

> A purposeful action is the more rational, the better the associated claim for efficiency can be justified. (Habermas, 1981, p. 27)

Whereas Habermas connects the rationality of a decision to the warrants for the underlying assumptions (pp. 26f.), our definition of rationality merely requires the strategy to be consistent with the agent's beliefs irrespective of whether the beliefs are warranted or not. This allows us to distinguish two questions: 'Is the decision rational?' and 'Is the decision founded on correct beliefs, i.e. are the underlying assumptions justified, are the *epistemic requirements* met?' Both questions are legitimate and important when evaluating alternative strategies and we might very well criticize a decision on the grounds that it rests on questionable beliefs (although, given these beliefs, it is perfectly rational).

Compared to these alternative notions of instrumental rationality, our definition can be termed a minimalist one. Adopting the minimalist definition in the course of the following investigations is justified for two reasons: First of all, rational decision making in the minimalist sense calls for an applicable bridge-principle, which is also a requirement of rational decision making in any stronger sense. Hence, if in some domain, as for instance economic policy-making, decisions cannot be taken according to the minimalist definition because no bridge-principle is available, rational decision making is not possible in any sense at all. Secondly, it shares the virtue of all minimalist definitions by allowing us to formulate more fine-grained questions. Accordingly, we can distinguish between a decision being rational (given some background knowledge) and its being based on justified beliefs.

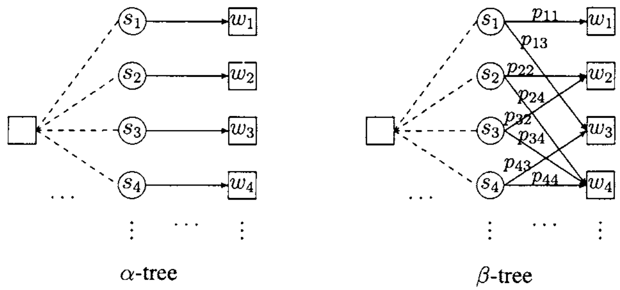



Figure 11.1: The $\alpha$- and the $\beta$-tree, representing epistemic situations which are required by the standard and the probabilistic pattern of rational conduct respectively.

## 11.2 The standard pattern of rational conduct

*The bridge-principle and its epistemic requirements*

The classical pattern of rational decision making is defined by the following bridge-principle:

**Bridge-principle 1 (Standard pattern)** *If strategy $s$ leads to the most desired world-state $w$ of all possible world-states that could result from some strategy, then the rational agent implements $s$.*

Consider as an illustration a decision by the Federal Open Market Committee to buy or sell federal funds in order to maintain interest rates at its target rate. In this example, the desired world-state is described by the target rate, and the alternative strategies are to buy or to sell federal funds, or not to intervene in the market. Given the knowledge of how the market intervention influences the interest rate, the bridge-principle BP1 unambiguously determines the strategy to chose.

The principle BP1 connects decisions on one side with beliefs about alternative strategies, their (certain) consequences, possible world-states and preferences between them on the other side. Hence, it makes the following epistemic requirements:

(i) The agent has to know the set of alternative strategies $S = \{s_1, s_2 \ldots\}$.

(ii) The agent has to know the set of possible world-states $W = \{w_1, w_2 \ldots\}$.

(iii) The agent has to have consistent preferences between the possible world-states so as to label one of them the most desired one. (An utility function might serve as a shortcut for such a preference relation.)

(iv) The agent has to know the conditional forecasts for all strategies (these can be summarized in a mapping from $S$ to $W$).

The left-hand graph in figure 11.1 represents the epistemic requirements of BP1. It will be referred to as $\alpha$-tree in the following.

If the standard pattern is applicable, that is all requirements (i) to (iv) are satisfied, the unique rational decision can be deduced.[3] The 'problem' can be 'solved', and rationality can hence fulfill its guiding function. Notice finally that the standard pattern requires normative judgment in the form of preferences between world-states. That is why the whole decision requires more than calculation and deduction, although, given the epistemic requirements, finding the rational decision is simply a technical task.

*Unfulfilled requirements in economic policy-making*

Are the epistemic requirements (i)-(iv) satisfied in the case of macroeconomic policy-making? I maintain that this is not the case. The reason is that we do not have credible macroeconomic forecasts conditional on alternative policy-strategies (requirement iv). This has yet to be shown for so far, i.e. in the first two parts of this study, we have just seen that and explained why categorical macro forecasts fail. However, in the course of the introduction I argued already that conditional forecasts are only credible if the corresponding categorical forecasts are so (see p. 11 above). As a consequence, conditional macroeconomic forecast are not credible either.

Although I consider the preceding argument as conclusive, I want to put forward a further reasoning that undermines the possibility of policy-relevant conditional macroeconomic forecasting. In order to avoid confusion, note that we are not talking about conditional forecasts in general but about very specific conditional forecasts — genuine policy conditional forecasts that predict what would happen were a specific policy-measure adopted. Now, consider the following methodological argument which draws an analogy: Genuine CFs are produced by the same methods that are used for categorical macroeconomic forecasting. But these methods largely fail to generate successful categorical forecasts. So, in analogy, macroeconomic CFs are (probably) incorrect, too. Eventually, all the explanations of forecast failure of part 2 directly apply to genuine conditional forecasts, too.[4]

However, the fact that positive macroeconomic CFs are not credible does not imply that this holds for negative conditional point forecasts, too. And indeed, that seems not to be the case. It is for instance certainly possible to correctly

[3] Notice that we have to make the implicit assumption that two different strategies do not lead to one and the same future world-state. This is quite reasonable, for even those strategies which seem to bring about the same consequences usually differ in some minor side effects they cause.

[4] This kind of argument will be vital for the reasoning in the next section.

predict that GDP will not grow by 12% if a government introduces an obligatory minimum holiday of two months for every employee. So, negative conditional macroeconomic forecasting appears to be quite feasible — but that does not safeguard the standard pattern of rational conduct.

We have seen that the standard pattern of rational conduct as defined by BP1 cannot be applied to macroeconomic policy-making because the epistemic requirement (iv) is not satisfied. In the literature on uncertainty, for example in the writings of Funtowicz and Ravetz that will be discussed in chapter 13, the absence of agreed-upon aims and values plays a role as important as the lack of credible CFs. What I want to demonstrate now is that this uncertainty about values might simply result from a lack of credible forecasts: The lack of correct CFs can always be reconstructed as an uncertainty about aims (condition (iii)-problem) and what appears to be a normative problem ("We do not know what we want.") sometimes really is a problem of limited foreknowledge.[5] To understand this, consider the above example of the Fed's monetary policy. Assume that we have no credible CFs that tell us how a change in interest rates influences inflation and growth. On the other hand, the Fed knows very well which open market interventions to perform in order to attain a certain interest rate. Now, if we consider the different open market interventions as alternative strategies and the possible future macroeconomic conditions as world-states, condition (iv) is not satisfied whereas condition (iii) is: The Fed aims at moderate, stable inflation and high growth rates. However, if we consider short-term ahead world-states instead which basically differ with respect to the interest rate, conditional forecasts that map the open market interventions into these world-states are available (condition iv) but the Fed does not have clear preferences between the world-states, hence condition (iii) is unsatisfied. In that second case, it seems as if the Fed merely has a 'normative' problem and must simply make up 'its mind' about what it wants. This, of course, is a fallacy; the apparent uncertainty about aims is deeper rooted in an uncertainty about consequences and outcomes. That is also true for the similar case of targeting a global CO2-concentration (see IPCC, 2001*a*, pp. 612ff.) as well as for many other examples. Generally, such a confusion might lead to an overestimation of the importance of uncertainty about values (which certainly exists to some extent).

Our conclusion that epistemic requirement (iv) of the standard pattern is not met seems to be challenged by the following subversive question: Are subjective beliefs about the consequences of alternative strategies not enough to apply BP1 and to make a rational decision? This is certainly possible and such a decision

5 An observation that will indeed be useful for the reconstruction of some procedures discussed in the next chapter.

would be rational according to our definition. However, we have seen earlier that a rational decision might still be criticized for being based on unjustified, questionable assumptions. And that is exactly the case if it is based on subjective beliefs. Thus we are fully entitled to criticize for instance a government which bases its policy decisions on wishful thinking instead of credible CFs.

Up to this point we rather imprecisely said that the standard pattern cannot be applied to macroeconomic policy-making in general. However, not virtually every macroeconomic policy-measure is necessarily affected by what we have said so far. We will complete this section by explicating which kind of macroeconomic policy-questions cannot be tackled by the standard approach.

First of all, any policy-measure that is meant to alter the economic output is concerned. For applying the standard pattern to such a problem requires conditional growth-forecasts. Hence, tax-cuts, for example, public investment programs or any other policy that is meant to boost the economy may not be dealt with as proposed by the standard pattern of rational decision making. Secondly, policies that are to influence price levels cannot be determined by the standard procedure. For these obviously require conditional inflation forecasts — which we do not have, either. This mainly poses a challenge for the central banks. Thirdly, many economic policy-measures which are not meant to alter the output *indirectly* rely on GDP-forecasts, since they, according to the standard pattern, require a conditional forecast which is in turn based on an categorical co-prediction of GDP. Consider for example fiscal policies which are supposed to consolidate the government budget. The conditional forecasts which predict tax-revenues conditional on alternative taxing-schemes or -rates seriously depend on categorical GDP-forecasts, and insofar as we do not have credible GDP-forecasts, we do not get credible conditional tax-revenue-forecasts either. For the same reason, poor performance of growth forecasts poses a problem to monetary policy, too. And even long-term structural reforms of the pension system, the public health system or the labour market cannot be managed according to the standard pattern as the corresponding conditional forecasts rely on categorical, sometimes even conditional co-predictions of GDP.

Finally, there might be some 'economic' policies which fall in neither of these categories and to which the standard pattern could be applied irrespective of macroeconomic forecasts' performance. Some types of environmental regulations or policies to foster innovation and development of new technologies might be of this kind. Of course, these policy-fields might be (and probably are) governed by other uncertainties so that the standard pattern cannot be applied there either. We have never claimed that uncertainties about economic growth and inflation are the only relevant ones for policy-making.

## 11.3 The probabilistic account of rational decision making

*The bridge-principle and its epistemic requirements*

If not according to the standard pattern, maybe rational decision making in macroeconomic policy areas is possible according to a probabilistic account. The standard bridge-principle applied in economics and rational choice theory in cases of uncertainty is,

**Bridge-principle 2 (Expected utility maximization)** *The rational agent adopts that strategy $s_i$ which maximizes expected utility.*

The expected utility of a strategy $s_i$ is defined as,

$$\mathrm{EU}(s_i) := p_{i1}\mathrm{U}(w_1) + p_{i2}\mathrm{U}(w_2) + \dots$$

with $p_{ij} = \mathrm{P}(w_j|s_i)$ denoting the probability that outcome $w_j$ results from *implementing* strategy $s_i$.[6] The utility function U() represents a shortcut for both the agent's preferences between different outcomes as well as his risk preferences (see Jaeger et al., 1998, pp. 153ff.).

Let us consider monetary policy as an example, again. Assume the Federal Open Market Committee had to decide whether to change its target rate or not. If utility is supposed to be a function of inflation and GDP, and the probabilities of all possible outcomes (combinations of GDP and inflation) are given for all strategies (target-rates), then the optimal strategy which maximizes expected utility can be calculated. If the decision problem is simple enough in terms of alternative choices and possible outcomes, the optimal decision might even be deduced non-quantitatively as the following example shall illustrate. In June 2003, the Fed justified a decision to decrease interest rates as follows,

> The Federal Open Market Committee decided today to lower its target for the federal funds rate by 25 basis points to 1 percent. [...]
>
> The Committee perceives that the upside and downside risks to the attainment of sustainable growth for the next few quarters are roughly equal. In contrast, the probability, though minor, of an unwelcome substantial fall in inflation exceeds that of a pickup in inflation from its already low level. On balance, the Committee believes that the latter concern is likely to predominate for the foreseeable future. [Fed Press Release, June 25, 2003]

6 Thus, what we deal with is for example not the probability that there are few malaria-deaths conditional on the fact that there are few blankets in the barracks, but the probability that there are few malaria-deaths conditional on the fact that the doctors burn the blankets.

In the rational reconstruction of the Fed's decision, utility will be considered as a function of inflation only.[7] Now, if we assume that

(a) there are three alternative strategies, namely to increase the target rate ($s_1$), not to change it ($s_2$), or to lower it ($s_3$);

(b) each strategy might result in a "pickup in inflation" ($w_1$), roughly stable inflation ($w_2$) or "an unwelcome substantial fall in inflation" ($w_3$) while there is a trade-off between pickup and fall; specifically we assume that,

$$\begin{aligned} \mathrm{P}(\mathrm{pickup}|s_1) + \mathrm{P}(\mathrm{fall}|s_1) &= \mathrm{P}(\mathrm{pickup}|s_2) + \mathrm{P}(\mathrm{fall}|s_2) \\ &= \mathrm{P}(\mathrm{pickup}|s_3) + \mathrm{P}(\mathrm{fall}|s_3); \end{aligned}$$

(c) the utility function satisfies $\mathrm{U}(\mathrm{fall}) < \mathrm{U}(\mathrm{pickup}) < \mathrm{U}(\mathrm{stable})$;

(d) the conditional probability of a fall in inflation is higher than the conditional probability of a pickup for all strategies whereas the difference between both probabilities is minimal for the strategy $s_3$;

it follows with BP2 that the rational agent adopts the third strategy.

According to the trade-off condition (b),

$$\begin{aligned} p_{31} + p_{33} &= p_{21} + p_{23} \\ \Leftrightarrow p_{31} - p_{21} &= p_{23} - p_{33}, \end{aligned}$$

where $p_{ij} := \mathrm{P}(w_j|s_i)$. With (c), this yields

$$\begin{aligned} (p_{31} - p_{21})\mathrm{U}(\mathrm{pickup}) &> (p_{23} - p_{33})\mathrm{U}(\mathrm{fall}) \\ \Leftrightarrow p_{31}\mathrm{U}(\mathrm{pickup}) + p_{33}\mathrm{U}(\mathrm{fall}) &> p_{21}U(\mathrm{pickup}) + p_{23}U(\mathrm{fall}). \end{aligned}$$

As $p_{32} = p_{22}$, it follows from the definition of expected utility that the expected utility from the third strategy is greater than the expected utility from the second one. A similar argument shows that the third strategy is also superior to the first one.

This exercise was not supposed to clarify the actual decision process of the Fed but to illustrate that a qualitative deduction of an optimal solution apparently requires strong assumptions and is therefore limited to sufficiently simple cases. Generally, numerical estimations of the specific probabilities are required for the calculation. But even in our simplistic case we probably cannot get along without such numerical estimations because they are what the assumptions (b) and (d) ultimately have to be based on.

7 Omitting the growth aim seems to be justified since "upside and downside risks" are judged to be "roughly equal".

The bridge-principle BP2 establishes logical connections between decisions on the one hand and beliefs about the possible outcomes of different strategies, the conditional probabilities of such outcomes and preferences between them on the other one. It makes the following epistemic requirements:

(i') The agent has to know the set of alternative strategies $S = \{s_1, s_2 \ldots\}$.

(ii') The agent has to know the set of possible world-states $W = \{w_1, w_2 \ldots\}$.

(iii') The agent has to have consistent preferences between the possible world-states as well as risk preferences. (An utility function might represent these preferences.)

(iv') The agent has to know the conditional density forecasts for all strategies, or, the other way around, the probabilities of the possible world-states conditional to the different strategies.

The $\beta$-tree in figure 11.1 represents the epistemic requirements of the probabilistic pattern of rational conduct. As for the standard pattern, the one and only rational decision can be calculated if the epistemic requirements are met. Its major advantage compared to the standard pattern is obviously that it no longer requires perfect foresight. On the opposite, it can be applied under uncertainty as long as reliable conditional probability forecasts are available. We will subsequently refer to such situations in which probabilities can be assigned to possible outcomes as decisions *under risk* in accordance with the terminology proposed by Knight (1921),

> It will appear that a *measurable* uncertainty, or 'risk' proper, as we shall use the term, is so far different from an *unmeasurable* one that it is not in effect an uncertainty at all. We shall accordingly restrict the term 'uncertainty' to cases of the non-quantitative type. (p. 20)

Let us finally notice how closely BP1 and BP2 are related: BP2 not only applies to cases where BP1 no longer applies, but it also applies to all cases where BP1 does. In these cases, BP2 reduces to BP1 so that the latter is in fact a special case of the former.

### *Unfulfilled requirements in economic policy-making*

Are the epistemic requirements of the probabilistic pattern fulfilled in the field of macroeconomic policy-making? Although these requirements are weaker than those of the standard pattern, they are not satisfied either. While the crucial requirement under certainty was (iv), the crucial requirement under risk becomes (iv'), namely the availability of credible conditional probability forecasts (CPF). In the preceding parts of this investigation, we have hardly dealt with probability forecasts so that an argument is needed which links the failure of macroeconomic point forecasts and their explanation to the impossibility of probabilistic forecasting. The following general claim is at the heart of my argument: The explanations of macroeconomic forecast-failure entail the failure of density forecasts. Almost all the problems CFs face and which were identified in part 2 are also virulent for CPFs, as we shall now see.

External effects (chapter 5). Many of the external effects an economy's development depends upon are not predictable. Even worse and what matters here, not even the probabilities of the possible outcomes are known. What is the probability that the civil war in X-land will end by Christmas? What is the probability that some scientific invention will revolutionize the way we produce Y-components? As these situations are anything but well-understood, shielded nomological machines and we have no probabilistic models to adequately describe them, we are not in a position to generate credible CPFs.

Data (chapter 6). Poor data quality does not necessarily represent a problem for probability forecasting: If GDP-data, for instance, were given as an objective density function rather than a point figure, point-forecasts would automatically evolve into density forecasts and thus give rise to PFs. But that is not the case. And we might doubt that it even could be the case: As we have seen, the statisticians themselves characterize the production of GDP-data as a rather creative process. This, however, means that any probability distribution quantifying the data imprecision would merely reflect the statisticians' subjective probabilities instead of objective probabilities.

Reflexiveness (chapter 7). No matter whether the published forecast is a probabilistic one or not, it enters into the expectation forming process of the economic agents as a new datum and therefore potentially alters its very object. Reflexiveness thus is as much an obstacle to probabilistic as to non-probabilistic forecasting.

Sensitive dependence on initial conditions (chapter 8). SDIC as such does not rule out probabilistic forecasting. If the probability distribution of the initial condition's error is given, we, or our computer, might very well calculate a density forecast. However, SDIC is like a lens which magnifies errors in the probability estimation. Assume that we had reliable probability estimations for most initial conditions except for some minor parameters. Maybe, they could have been neglected if the system were linear, but under SDIC, these minor parameters might be crucial and the shape of the forecasted density function might very much depend on their probability distribution. This is why SDIC magnifies all the obstacles discussed in part 2. In addition, the fundamental limit of predictability in terms of a maximum forecast horizon derived in section 8.1 applies (primarily, indeed) to probability forecasting as well.[8]

Simulation (chapter 9). As we have explicitly discussed, probability forecasting which makes use of simulations requires the availability of a reliable simulation. We do not have such reliable economic or climate simulations at our disposal and the simulations actually performed do not give rise to credible probability forecasts.

Concluding the previous paragraphs, the explanations of macroeconomic forecast failure imply a failure of conditional probability forecasts, too. Hence, the epistemic requirement (iv') of the probabilistic pattern of rational conduct is not satisfied in the field of macroeconomic policy-making and the principle BP2 cannot be applied.

In analogy to the non-probabilistic case, this does not exclude the possibility of negative probability forecasting. More specifically, if some negative point forecasts are credible, so are some negative probability forecasts, too.[9] For every negative point forecast implies a negative density forecast, namely the forecast which rules out any density function which attributes non-zero probability to the event excluded by the negative point forecast. Thus, negative probability forecasting is possible to the degree that we can rule out density functions which attribute non-zero probabilities to impossible outcomes. On the other hand, these are the only density functions we can rule out. The arguments against probabilistic forecasting imply that we can forecast the objective probabilities of the *possible* outcomes neither positively nor negatively in a credible way.

A critical question one faces at this point of the argumentation is: 'Very well, you've shown that we won't get objective probability forecasts. But aren't subjective probabilities sufficient to apply (Bayesian) decision theory?' Subjective probabilities are estimates of probabilities which are not necessarily justified in any way but which merely represent an individual's subjective degrees of belief. The answer thus is similar to the one we gave in the previous section: Of course, BP2 and the whole apparatus of EU maximization can be applied if subjective CPFs are given. And by doing so, one might derive the one and only rational de-

---

8 Hitherto, I have assumed that probability forecasts are derived from a model with a deterministic dynamic and initial probability distributions of the variables. This is in fact part of the procedure by which the Bank of England constructs its fan charts (see Britton et al., 1998). Still, probabilities are also used to characterize a fully deterministic dynamic system. Assuming that it is ergodic and analyzing its attractor enables one to derive the probability that the system is in a certain subspace of the phase diagram at an arbitrary time $t*$. But the crucial notion here is that of arbitrary time for this method yields objective probabilities only if the time $t*$ is chosen randomly by a well-understood random mechanism from a sufficiently large time interval. Objective probabilities are introduced by random sampling only. If one attempts to forecast the system state at $t'$, $t'$ is, however, not chosen randomly. For a similar argument compare Gillies (2000, p. 192)

9 If we consider a point forecast as equivalent to a very narrow interval forecast.

cision. But rationality is, to say it again, not the only ground on which to criticize a decision. Knowing how well-founded the beliefs and assumptions underlying a decision are is as important as the question of its rationality. And in the case of subjective probability estimations, we have to ask why collectively binding decisions should be based on some set of consistent probabilistic beliefs of some individual if these beliefs lack any justification? Why, then, these and not any other beliefs? No, if we do not know the density functions, if we do not have credible CPFs, then there is no point in taking some subjective beliefs for granted and *pretending* to possess the required CPFs. That would be (self-)deception, and would have nothing to do with responsible decision making.

## 11.4 A causal account of rational conduct

The notion of conditional forecast has been one of the key concepts in the patterns of rational conduct we have discussed so far. Maybe we have to change our conceptual perspective in order to find a pattern which can be applied in economic policy-making. Maybe the notion of conditional forecast is not at the heart of rational decision making at all. If we succeed in setting up a bridge-principle that does not comprehend credible conditional forecasts as one of its epistemic requirements, then rational decision making would be unaffected by poor forecast performance. But what could replace the central notion of conditional forecast? An obvious candidate is the idea of causality. A bridge-principle which links decisions with beliefs about facts and preferences by exploiting cause-effect relationships is

**Bridge-principle 3 (Cause-effect)** *If strategy $s_i$ causes the desired world-state $w$, then the rational agent implements $s_i$.*

This principle is easily illustrated: Assume the Bank of Japan (BoJ) knew that a lower interest rate would cause the economic agents to realize more investment projects which in turn would boost economic growth. And assume that this were the most-desired thing the BoJ could think of. Then, in line with BP3, the rational decision of the BoJ would be to decrease interest rates.

At first glance, the epistemic requirements of BP3 seem to be less demanding than those of the previous bridge-principle,

(i") The agent has to know the set of possible world-states $W = \{w_1, w_2 \ldots\}$.

(ii") The agent has to have preferences between the possible world-states which determine a most desired state $w^*$.

(iii") The agent has to know which strategy $s_i$ causes $w^*$.

In short: The agent must know what he wants and which intervention causes what he wants. Although this seems to be satisfiable, I am going to argue that the requirements (i")-(iii") are in fact not sufficiently weaker than those of bridge-principle 2. And insofar as the latter cannot be met in the domain of economic policy-making, the former cannot either. I see two kinds of arguments which implement this general line of reasoning; one of them is going to be fully elaborated now, whereas I can merely sketch the other one.

Here comes the first argument. Its main idea consists in showing that any decision making relevant causal claim implies negative CPFs of the kind we cannot make in a credible way. Assume that conditions (i") to (iii") were satisfied. It is analytically true that the probability of $w^*$ given $s_i$ is either larger or strictly less than the probability of $w^*$ given that $s_i$ is not carried out. Now, if it is strictly less, then following the cause-effect principle of rational conduct would lead us to adopt a decision which actually decreases the probability of the desired outcome! That does not only sound intuitively absurd but also contradicts the principles BP1 and BP2. Hence, in such a case, BP3 is not suited for economic policy-making. But what if the probability of $w^*$ is increased by implementing $s_i$? If we knew reliably that that were the case, we would be in a position to derive correct negative conditional density forecasts. In particular, we would be able to rule out all density forecasts that do not satisfy the above probability inequality. These, however, would clearly include density functions that assign zero-probabilities to all impossible outcomes, density functions that we cannot credibly exclude according to the discussion in the previous section. Hence, we do not have decision making relevant causal knowledge with regard to economic policy-making. The argument is in brief,

(1) (i"), (ii") and (iii").
(2) Either $\mathrm{P}(w^*|s_i) \geq \mathrm{P}(w^*|\neg s_i)$ or $\mathrm{P}(w^*|s_i) < \mathrm{P}(w^*|\neg s_i)$.
(3) If $\mathrm{P}(w^*|s_i) < \mathrm{P}(w^*|\neg s_i)$, then BP3 implies to chose a strategy which decreases the probability of the desired outcome.
(4) If some principle implies to chose strategies that decrease the probability of the desired outcome, then this principle is not suited for rational decision making.
(5) The probability inequality $\mathrm{P}(w^*|s_i) \geq \mathrm{P}(w^*|\neg s_i)$ implies some negative conditional forecasts of density functions which assign zero probabilities to all impossible outcomes.
(6) *Thus:* If we have appropriate causal knowledge so as to apply BP3, then we are in a position to establish negative conditional forecasts of density functions which assign zero probabilities to all impossible outcomes.
(7) In the case of economic policy-making, we are not in a position to establish such negative conditional density forecasts.

(8) *Thus:* We do not have appropriate causal knowledge so as to apply BP3 in the field of economic policy-making.

This argument makes it clear that the original intention to replace the notion of conditional forecast by the notion of causality fails. Causality is the stronger concept, and relevant causal knowledge implies some knowledge about negative conditional probability forecasts we just do not have.[10]

A further, second argument which concludes that (iii") is satisfiable in economic policy-making would consist in showing that the methods which are used to derive causal claims in economics are in fact undermined by the very same problems that hamper macroeconomic forecasting. Although I believe that this is the case, I am not able to demonstrate it. Such a task would lead us beyond the scope of this inquiry.

## 11.5 Uncertainty and ignorance

Our macroeconomic foreknowledge is neither adequately represented by the $\alpha$-tree nor by the $\beta$-tree. That is the reason why the standard as well as the probabilistic pattern of rational conduct cannot be applied. On the other hand, we are certainly not in a situation of complete ignorance. We definitely know more about our economic future, e.g. things that will not happen like full-employment by next week, than a stone-age man who came here traveling through time. The problem sparked off by the collapse of the traditional bridge-principles is that any decision, any strategy is consistent with this background knowledge. The idea of rationality can no longer fulfill its normative function. And this situation will prevail unless we find new bridge-principles that fully take account of the limits of our (economic) foreknowledge and do not make epistemic requirements that are not satisfiable. Or, as the IPCC wisely puts it in its (as we will see: twofold) conclusion,

> In sum, a strategy must recognize what is possible. (IPCC, 2001*b*, p. 774)

The presentation and discussion of such new bridge-principles is what the next chapter is dedicated to. Yet before we can propose bridge-principles whose epistemic requirements can be met, we have to specify our epistemic situation. What

[10] Some readers might wonder why I included the distinction in premiss (2) with this strange second case where the realization of a cause decreases the probability of its effect. Isn't this a mere logical possibility only philosophers bother about? Philosophers bother about such possibilities indeed and philosophers of science have in fact invented ingenious devices, experimental settings (see for instance Cartwright, 1999, p. 163) where a cause is claimed (not unobjected) to decrease the probability of its effect. That is why I included the case distinction.

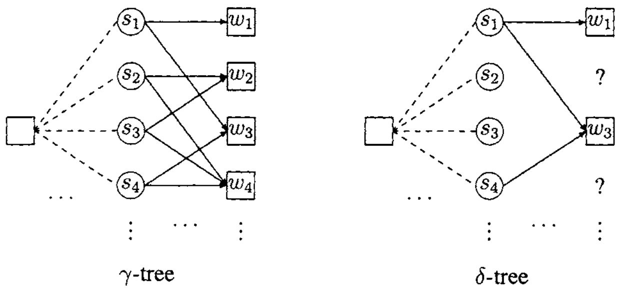



Figure 11.2: The $\gamma$- and the $\delta$-tree, representing the epistemic situations of uncertainty and ignorance.

kind of foreknowledge do we have if we have less than represented by the $\beta$-tree and more than nothing?

So far, we have dealt with two kinds of epistemic situations: those of certainty and risk. Now, we will add two more categories: situations of uncertainty and of ignorance. In line with Knight's original proposal, uncertainty refers to situations where one knows the possible outcomes of alternative strategies without being able to assign probabilities to the resulting world-states. In short, not *probabilities* are known but merely *possibilities*. Under ignorance, things are even worse: Here, we do not even know all the possible outcomes of an action. When, for example, in the 1930s engineers and chemists invented a new kind of substance with stable thermodynamic properties and desired safety characteristics that could be used in a variety of industrial processes as well as in households, they did not even dream of the possibility that these chlorofluorocarbons (CFCs) might trigger off the depletion of the stratospheric ozone layer. Under ignorance, we might know some possible outcomes and we might know some impossible outcomes, i.e. what will not happen, but there are also some possibilities of which we are unaware. This implies that there can be different degrees of ignorance which we cannot estimate (for we cannot count possibilities we have not thought of, yet). The epistemic situations of uncertainty and ignorance are represented by the $\gamma$- and the $\delta$-tree in figure 11.2.

Our simple four-fold category-system by and large agrees with more sophisticated systems which have been proposed by Faber et al. (1996), Wynne (1993) or Healy (1999). It is sophisticated enough to guide our further investigations. In particular, our account can be reconciled with Levi's interesting extension of traditional rational choice theory. Levi (1980) allows for multiple *permissible* probability- and utility functions to account for uncertainty. This enables him to investigate whether a given strategy is *E-admissible*, i.e. maximizes expected utility for some combination of the different probability- and utility functions (Levi, 1980, p. 96). If there is a single *E-admissible* strategy, then this is the optimal choice according to Levi. Now, given that we cannot rule out any density function that assigns zero probabilities to impossible outcomes, we have to deal with situations where any such probability function is *permissible*. This implies in Levi's theory that almost[11] any strategy is *E-admissible* and that the agent has to rely purely on secondary criteria. Bridge-principles for decision making under uncertainty are nothing but explications of such secondary criteria.

As far as our macroeconomic foreknowledge is concerned, we are in situations of uncertainty or more or less severe ignorance, depending on the specific policy-problem. The task of the next chapter will therefore be to develop bridge-principles that are applicable under such conditions.

Some of the bridge-principles and their patterns of rational conduct discussed in the next chapter rely on the assumption that uncertainty or ignorance will eventually be reduced to risk or certainty. It would be a crucial set back for such patterns of rational conduct if some kind of uncertainty or ignorance were irreducible, indeed. The possibilities of improving one's foreknowledge and thereby shifting step by step from ignorance to risk shall therefore be discussed in the remainder of this chapter. Two reductions have to be considered: (a) the reduction of ignorance to uncertainty, and (b) the reduction of uncertainty to risk.

I reckon step (b) as the by far more important obstacle whereas Faber et al. (1996) argue that there is even irreducible ignorance. So let us consider step (a) first: Is there irreducible ignorance? First of all, I take it that whatever is physically possible can be imagined with some creative effort. There is no possible world that we cannot imagine in principle. In this sense, ignorance can be reduced. However, this is rather a reduction in degree, and it does not follow that we can actually imagine all possible outcomes, i.e. reduce ignorance to uncertainty.[12] Furthermore, there might be empirical reasons that prevent the complete reduction of ignorance to uncertainty: We simply do not have enough time — let alone cognitive capacities — to enumerate all possible outcomes. Thus, as a matter of fact, we will not always succeed in reducing ignorance and lots of decision situations are characterized by ignorance. In addition, we can never be sure that we have eliminated ignorance unless the situation is sufficiently simple so that we merely have to consider a handful of quantitative variables whose combinations make up the possible world-states.

Although there are some empirical obstacles to the reduction of ignorance, there are apparently no fundamental limits. This is yet the claim Faber et al. (1996) try to defend. They put forward five arguments which we can easily refute. The

[11] To be specified in section 12.2.

[12] Although some paintings might be imitations, it is not possible that all paintings are imitations.

first two arguments fail because of a confusion of unpredictability and ignorance. Even if the problems of "genotypic evolution", i.e. a change in a system's fundamental structure, and "chaos" imply that the system's behavior is unpredictable, as Faber et al. (1996, pp. 217f.) point out, this does by no means entail that we cannot "specify all possible future outcomes" (Faber et al., 1996, p. 210). The next three arguments all are examples of poor philosophy of science. Consider this,

> So scientists have to accept that, despite their sharpest definitions, they have to use words and notions which are not completely unambiguous. From this it follows that scientific statements also can never be totally clear and unambiguous. So we remain ignorant even if we express our surest knowledge [...].

However, ambiguity is no obstacle to describing possible future world-states that might result from our strategies, and is thus no obstacle to reducing ignorance. The fourth argument starts by claiming that the "ideal scientist can derive the entire corpus of his knowledge from [...] an axiomatic system" (p. 219). But "we can never know the truth of our axioms". (p. 220) That is why we "remain ignorant at a very fundamental level of our scientific endeavor" (p. 220). Sure, no empirical theory can be proven to be true. But what does this have to do with imagining possible worlds? Nothing, as we do not need 'really true' theories to think of what might happen if ... The most fantastic argument, however, is the fifth one: Its main idea is in short that from "Gödels theorem we [...] know that even closed logical systems are sources of ignorance, and this ignorance cannot in principle be reduced" (p. 220). This clearly takes the cake! The (non-)provability of some sentences in some formal systems has definitively no impact at all on our ability to reduce ignorance by using our fantasy! But in some sense, finally, Faber et al. (1996) are right: Ambiguity is a problem. It is the problem which underlies all their fallacious arguments as they do not use the notion of ignorance in the way they, like we, originally defined it.

So, let us have a closer look at step (b), the reduction of uncertainty to risk. Here, we should distinguish (b1) a reduction as the 'forecast horizon' decreases and new information emerges and (b2) an immediate reduction which consists in constructing the $\beta$- out of the $\gamma$-tree. In so far as (b1) is concerned, reduction of uncertainty due to newly available information seems to be a plausible idea. Observation 4 from part 1 which stated that macroeconomic forecast error typically increases with the forecast horizon supports this intuition. That much is clearly true: Uncertainty is reducible in the sense that, based upon new information, some world-states which were formerly thought to be possible may be ruled out so that the fan of possible future world-states narrows.[13] This, however, does not mean that we shift from one epistemic category to another: Still, we face uncertainty, not risk. Uncertainty is reduced in degree, but it is not reduced to risk. And as the reasons why we do not obtain any credible probability forecasts are invariant to the forecast horizon, I do not really see how such a shift could occur.[14]

This said, it seems to be unlikely that there is a way to construct the $\beta$- right out of the $\gamma$-tree (b2). We have considered the categories of uncertainty and ignorance only because neither the notion of risk nor certainty adequately described our epistemic situation, our (economic) foreknowledge. Insofar as uncertainty is a weaker requirement than risk, how could it be possible to deduce a $\beta$- from a $\gamma$-tree? There is no such deduction. Nonetheless, I will briefly discuss what might look like one. Kahnemann et al. (1982), based on experiments about probabilistic inference-making, identify three heuristics which individuals tend to rely on when making probabilistic judgments. The first of these heuristics, representativeness, consists in estimating the probability that $a$ is $G$ according to the similarity of $a$ to some typical instance of $G$. The second heuristic, availability, suggests to determine the probability of some state $p$ by the difficulty to imagine that $p$. And following the third method, adjustment and anchoring, the probability of some state $p$ is adjusted from the presumably known probability of some state $p^*$ according to the degree of difference between $p$ and $p^*$. These heuristics and especially the second one could be used to construct the $\beta$- out of the $\gamma$-tree. But should they be used to do so? Not without pointing out their many pitfalls, Kahnemann et al. (1982) claim that these heuristics tend to be effective. I am sure that they are effective as long as they are applied in everyday situations the individual is acquainted to and experienced with. Yet they are not suited to assess probabilities of future states in highly complex systems — as long as these probabilities are meant to be objective (and that is what we suppose them to be). One of the reasons why individuals stick to such simple heuristics is that they enable them to make quick judgments without taking the complexities fully into account. But that is exactly what we ought to do in climate or macroeconomic policy: recognize the systems' complexity.

Our critique of attempts to reintroduce probabilities is backed by recent developments in operational research. In the introduction to *Rational Analysis for a*

[13] Yet, if we were in a position of ignorance, new information might on the contrary make us realize that there are more possibilities than we originally thought of.

[14] These reasons of forecast failure might no longer hold if science — economics or climatology — is revolutionized at some time in the future and equipped with a new ingenious method to generate forecasts. This changes the name of the game, and such a scientific progress is of course possible, and unpredictable. The irreducibility of uncertainty to risk in the course of time is therefore as firmly but not any firmer established as the meta-prediction that forecasts will fail in the future for the same reasons they fail today.

*Problematic World Revisited*, the standard textbook of 'soft' operation research, Rosenhead & Mingers (2001) point out that,

> where the uncertainties (about what is currently happening, about future events, about what priorities to apply) are an integral part of the 'wickedness' of the situation, attempts to reduce them to numbers can only result in the analysis being unhelpful. [...] When an eventuality whose occurrence is uncertain is important enough to be considered in deliberations, it is that *possibility* that is relevant, more than any numerical expression of its *probability*. (pp. 16f.)

Thus, again, the urgency to develop patterns of rational conduct for situations of uncertainty and ignorance.

# Chapter 12

# Rational decision making under uncertainty and ignorance


## Summary

This chapter explores alternative approaches to decision making under uncertainty and ignorance: the quasi-probabilistic approach, extremum approaches and approaches focusing on the spectrum of future options and possibilities. It defines the approaches by their respective bridge-principles before presenting sound applications and discussing their limitations. Also, it introduces the theory of sequential decision making which allows to apply the findings regarding one step decisions to sequential cases, too. In sum, this chapter demonstrates that there is no unique, entirely convincing approach to decision making under uncertainty and ignorance.


## 12.1 How to deal with ignorance

To which kind of belief do bridge-principles have to connect choices if our background knowledge is characterized by uncertainty and ignorance? Ignorance merely indicates the incompleteness of this background knowledge: The $\delta$-tree is nothing but a holey $\gamma$-tree. In other words, ignorance does not positively specify the kind of knowledge one possesses. For, in addition, it is impossible to estimate the degree of ignorance, bridge-principles for uncertainty and ignorance have to exploit the background knowledge's structure as represented by the $\gamma$-tree. We have to find principles which relate beliefs about conditional possibilities and conditional impossibilities to choices.

Although the principles we consider are, in the first place, principles which help us to deal with uncertainty, we can test in how far they are suited for situations of ignorance by examining their robustness to ignorance. Such a robustness test in a first step determines the correct choice according to a certain bridge-principle for some situation of uncertainty. Then, by deleting some of the assumed background knowledge, we introduce ignorance. The question is: Does the application of the bridge-principle to the modified background knowledge still give rise to the same choice? If so, then the principle obviously prescribes the 'correct' choice even if background information is limited, i.e. it is robust to ignorance. On the other hand, the more the prescribed choice varies under ignorance as compared to uncertainty, the less the principle is able to cope with ignorance.

But maybe, from a more practical point of view, the most important thing when making decisions under ignorance is simply to be aware of it. To be aware that something completely unforeseen might happen. And this awareness should result in being prepared. Such a — only apparently paradoxical — preparation for the unknown might consist in very different measures like saving some resources or increasing one's flexibility to react. For ignorance can always be reduced in degree, as we have argued in the previous chapter, one should in addition always try to identify omitted possibilities and improve one's background knowledge. As we have already noted, this rather involves creativity than rigid proof or systematic investigation. Thus, no recipes exist for conducting such an ignorance-reduction. In the light of this task, even the quite absurd methods of ensemble predictions in climatology which consist in constructing possible worlds become very useful. Finally, the development of comprehensive scenarios, which is also the approach adopted by the IPCC, might help not to forget some possibilities while, in the same time, rendering the information of the $\gamma$-tree in a way that does not overstrain human cognitive capacities (see also Chermack, 2003).

## 12.2 The quasi-probabilistic approach

*The mean-principle*

The first bridge-principle for decision making under uncertainty we discuss ranks the alternative strategies according to the mean-utility of their respective possible outcomes. Hence,

**Bridge-principle 4 (Mean-principle)** *The rational agent chooses a strategy $s$ so as to maximize the mean-utility of the possible outcomes.*

This principle is logically equivalent to BP2 plus the Principle of Indifference which says to assign equal probabilities to its possible outcomes for each strategy.[1] An agent who adopts the mean-principle therefore acts *as if* maximizing expected

[1] More precisely, the Principle of Indifference as originally stated by Keynes (1921, p. 42) asserts that equally justified alternatives are equally probable.

utility under the assumption of equal probabilities. Although the probabilistic approach cannot be applied under uncertainty (section 11.3), this as such is no argument against the mean-principle because for *any* choice an agent makes under uncertainty there are conditional probability distributions $\mathrm{P}(w_j|s_i)$ so that the agent acts *as if* maximizing EU given $\mathrm{P}(w_j|s_i)$. Thus, if the attribution of equal probabilities were justified, the mean-principle would be justified, too; whereas the inverse does not necessarily hold: Justifying the equal-probabilities attribution is not necessary for justifying the mean-principle. The mean-principle might eventually be justified for other reasons. But which ones? What, if anything, makes the mean-principle attractive? Its underpinning idea seems to be that all possibilities should be treated equally — not in terms of their evaluation, but in terms of their chance of occurrence. If for example five different outcomes are possible and we do not know their probabilities, why assume that one is more likely than another? But if the underlying reasons of the mean-principle are of this kind, then it is nothing but the maximizing EU principle given subjective probabilities indeed and there are no other reasons in favor of it. The mean approach is quasi-probabilistic, a short-cut for a special case of expected utility maximization. The argument against the mean-principle is in brief,

(1) If it is justified to assign equal probabilities to the outcomes of each strategy rather than assuming any other probability distribution, the mean approach is justified, too.
(2) This is the only justification of the mean-principle (mean approach is quasi-probabilistic).
(3) *Thus:* If and only if it were justified to assign equal probabilities to the outcomes of each strategy rather than assuming any other probability distribution, the mean approach would be justified.
(4) It is not justified to assign equal probabilities to the outcomes of each strategy rather than assuming any other probability distribution.
(5) *Thus:* The mean-principle is not justified.

In addition to the quasi-probabilistic character of the mean-principle, here is another problem to consider: The mean-principle is very susceptible to ignorance. The introduction of even small degrees of ignorance into the background knowledge might alter the choice entailed by the principle significantly. As an illustration consider the situation described in table 12.1. The agent has to choose between two alternative strategies $s_1$ and $s_2$. Under ignorance, strategies $s_1$ and $s_2$ seem to have four and three possible outcomes respectively with the evaluations given in 12.1a. As the mean utility of $s_2$ is larger than that of $s_1$, the mean-principle implies to implement $s_2$. Now, assume that the third possibility of $s_1$

(a) Ignorance

| strategy | utilities of possible outcomes | mean utility |
|---|---|---|
| $s_1$ | $1, 1, 2, -10$ | $-3/2$ |
| $s_2$ | $1, 2, -5$ | $-2/3$ |

(b) Uncertainty

| strategy | utilities of possible outcomes | mean utility |
|---|---|---|
| $s_1$ | $1, 1, 1, 1, 3, -10$ | $-1/2$ |
| $s_2$ | $1, 2, -5$ | $-2/3$ |

Table 12.1: The mean approach is not robust to ignorance.

which was evaluated with utility 2 in fact 'hides' three different possibilities with utilities 1, 1 and 3 which become apparent if ignorance is resolved (figure 12.1b). Under uncertainty, the new mean utility of $s_1$ is -1/2 and therefore greater than $s_2$'s mean utility so that the rational choice is $s_1$ instead of $s_2$. The main reason why the mean-principle reacts so sensitively to ignorance is that any information encoded in the $\gamma$-tree, in particular including the number of possible outcomes of each strategy, is used to determine a ranking.

Over and above these rather abstract reasons against the mean approach, there are numerous examples where the application of the mean-principle leads to counter-intuitive results. As these examples coincide with the positive examples in favor of a worst-case scenario approach, they will be dealt with in the next section.

### *Ruling out dominated strategies*

We have argued in the previous chapter that under uncertainty it is as reasonable to assume some set of probability distributions as any other.[2] This is the reason why the mean approach cannot be justified. However, it does not imply that every strategy can be presented as rational just by assuming an appropriate set of probability distributions. For some strategies might not maximize EU under any set of probability distributions. These strategies can consequently be ruled out under a probabilistic approach without presupposing objective CPFs. Under which conditions is that the case? Levi's extension of rational choice theory, briefly introduced in the previous chapter, as applied to the special case where the $\gamma$-tree and the utility-function are fixed, represents a suited framework to address this question.

Indeed, there is a single sufficient and necessary condition for being admissible (rational under some set of CPFs) if any set of probability distributions is permissible: A strategy $s_i$ is admissible if and only if for every strategy $s_j$ ($j \neq i$) there is a possible consequence $w_{k_i}$ of $s_i$ whose utility is greater than that of some

2 As long as they assign zero-probabilities to all impossible outcomes.

possible outcome $w_{k_j}$ of $s_j$. Formally,

$$\forall i(\forall j \neq i \exists k_i, k_j : U(w_{k_i}) \geq U(w_{k_j}) \iff s_i \text{ is admissible}).$$

The proof comes in two steps. But first of all, let us fix the following notation: $w_i^+$ shall denote the best possible outcome of some strategy $s_i$ and, accordingly, $w_i^-$ the worst one. Besides, the double index $j_i$ in "$w_{j_i}$" ranges over all world-states which are possible outcomes of $s_i$.

(a) To prove the first (left-pointing) implication, let $s_i$ be some strategy and assume that $\forall j \neq i \exists k_i, k_j : U(w_{k_i}) \geq U(w_{k_j})$. This clearly implies that $\forall j \neq i : U(w_i^+) \geq U(w_j^-)$. We then consider the following conditional probabilities,

(i) for $i$:

$$\mathrm{P}(w_k|s_i) = \begin{cases} 1 & w_k = w_i^+ \\ 0 & \text{otherwise} \end{cases};$$

(ii) for $\forall j \neq i$:

$$\mathrm{P}(w_k|s_j) = \begin{cases} 1 & w_k = w_j^- \\ 0 & \text{otherwise} \end{cases}.$$

If we compare the expected utilities of $s_i$ and some other strategy $s_j$ we thus get,

$$\mathrm{EU}(s_i) = U(w_i^+) \geq U(w_j^-) = \mathrm{EU}(s_j).$$

Hence, $s_i$ is by definition (see page 192 above) admissible.

(b) The second part of the proof consists in showing the contraposition of the left-pointing implication. Accordingly, let us assume that $\exists j \neq i \forall k_i, k_j : U(w_{k_i}) < U(w_{k_j})$. This entails in particular that $U(w_i^+) < U(w_j^-)$, i.e. the worst possible outcome of $s_j$ is still better than the best possible outcome of $s_i$. Yet this implies that whatever the weights (probabilities), the mean weighted utility (expected utility) of $s_i$ is strictly smaller than that of $s_j$. Hence, $s_i$ is not admissible.

What we have just proved in turn entails that a strategy $s_i$ is not admissible if and only if there is a strategy $s_j$ so that every possible outcome of $s_j$ has a larger utility than any of $s_i$'s possible consequences. In other words, $s_i$ is not admissible if and only if $s_j$ dominates $s_i$.[3] It is definitively reasonable to rule out dominated strategies, but it does not help us very much because strategies in the decision problems we are interested in are *de facto* hardly ever dominated. Even if we are lucky and some policy-measures can be ruled out by this criteria, there are usually a lot of strategies left. And the case where one strategy dominates all alternatives is virtually never realized. Therefore, ruling out dominated strategies is not a bridge-principle which could fulfill its guiding function in macroeconomic policy-making.

3 As Levi (1980, p. 144) notes incidentally, too.

## 12.3 Extremum principles

The susceptibility to ignorance of the mean approach arose because it exploited any available information, that is the entire $\gamma$-tree. If bridge-principles attempt to avoid this, they have to pick some particular parts out of the $\gamma$-tree and base the evaluation of alternative strategies on these specific parts only. A straightforward way to do so consists in considering only a specific subset of possible consequences for each strategy. And the possible consequences which suggest themselves are, of course, the extreme outcomes, i.e. the best and the worst possible case. Especially the latter has been used extensively to formulate bridge-principles for uncertainty and ignorance.

*The bridge-principle and alternative theoretical formulations*

The best known extremum bridge-principle is the so-called maximin rule which states that one should try to maximize the utility of the worst possible consequence,

**Bridge-principle 5 (Maximin)** *The rational agent chooses the strategy with the most acceptable worst possible outcome.*

If there are two strategies with the same smallest possible utility, then the second worst possible outcome is considered and, if necessary, the third etc., etc. to establish the ranking between the two strategies. This is an application of the "lexicographical rule"; and the maximin principle, when applied together with this rule, is also called "leximin".

Maximin has been discussed in many different versions in decision theory (see Levi, 1980, pp. 145ff.): Defining regret as the difference in utility between the actual outcome and some fixed best possible outcome, minimax regret is nothing but the maximin rule.[4] The same holds for the minimax risk (Luce and Raiffa) or minimax loss (Savage) rules.

A more general approach which considers both extreme possible outcomes is the optimism-pessimism principle originally proposed by Shackle (1949). A simplified version according to Levi (1980, p. 148) is this[5]: After a focal pair consisting of its best and worst possible outcome is attributed to each strategy, the ranking of the strategies is determined by some procedure which ranks the focal pairs. The optimism-pessimism approach does not favor or prescribe a specific ranking method of focal pairs. This is left to the decision maker. One family of

---

4 If, however, maximal regret were defined as the difference between best and worst possible outcome for each action (Hansson, 1997, pp. 297f.), it would lead to clearly absurd consequences as for instance adopting options with very, very bad outcomes only.

5 A more detailed, but still brief exposition can be found in Ford (1990).

ranking methods for focal pairs is the weighted average ranking $\alpha U(w^-) + (1 - \alpha)U(w^+)$. Given this ranking procedure, the value of $\alpha$ depends on what Shackle calls the agent's "gambler preferences" (Ford, 1990, pp. 28f.). Thus, a security oriented agent might choose a value of $\alpha = 1$ so that the optimism-pessimism rule reduces to the maximin principle, whereas a more risk-seeking individual or an incorrigible optimist would probably chose a much smaller value for $\alpha$. What is important about the optimism-pessimism approach is that it is a unifying frame for similar principles and not a bridge-principle as such. If adopted by an agent, it does not prescribe any choice in a decision making situation. Though it helps to focus the choice on two instead of possibly infinitely many possible outcomes, it hardly fulfills a bridge-principle's guiding function.

*Applying maximin*

Many apparently different approaches to deal with uncertainty and ignorance which have been developed by practitioners facing real decision problems reduce to the maximin principle at a second glance. There are consequently at least as many different practical versions as there are theoretical formulations of the maximin approach.

Before, however, we proceed to the practical applications, we should mention the famous application of the maximin rule in political philosophy by John Rawls. Rawls (1971) suggested that individuals under the veil of ignorance apply the maximin principle. This was attacked by Bayesians like Harsanyi (1975) but has recently been defended by Angner (2004) on the basis of Levi's theory and reasonings which are very similar to ours. These parallels are less surprising if one realizes that Rawl's fictitious veil of ignorance is not a fiction at all according to the results of our investigation — at least not in so far as our economic foreknowledge is concerned.

We are all familiar with maximin principles as we apply them in our everyday reasonings. Although the "maximin-rule" is probably hardly known to decision makers, the notion of "worst-case scenarios" is certainly a familiar one. But worst-case reasoning is of course nothing but maximin reasoning. Consider the following situation: Having got the flue you consult Dr. Newmed. He gives you the advice to take one of the new WonderPills that will possibly cure you at once so that you won't have to bother with the flue for the next two or three weeks. After you asked him what he means exactly with "possibly", he reluctantly admits that these pills are his own invention and nobody had ever tried them before. Given his theory, there is the realistic possibility that they cure at once, but he cannot rule out any "side-effects", especially, in the *worst case*, irreversible damage to the brain. This said, you say good bye and go. And in doing so, you apply the maximin principle for decision making under uncertainty and ignorance.

Another concept we are quite familiar with and that frequently appears in political discussions is the notion of robustness. If we do not know what is going to happen, whatever we do should be robust to the unpredictable developments. This means that a decision should ensure the attainment of goals irrespective of what is going to happen.[6] Yet if a strategy *guarantees* the attainment of some aims, even the worst possible outcome of that strategy is still acceptable. On the other hand, if some strategy has the most acceptable worst possible outcome relative to all alternative strategies, then this strategy is more robust and guarantees the attainment of some goals to a higher degree than any other strategy. Hence, focusing on and maximizing the worst possible outcome is a short-cut procedure for ensuring robustness.

One of the very few approaches suited for decision making under uncertainty discussed in the third volume of the IPCC's Third Assessment Report, *Mitigation*, is the tolerable windows approach (see IPCC, 2001*a*, pp. 616f.).[7] This approach consists in identifying those strategies which ensure that the system evolves within a predefined corridor of acceptable (tolerable) states. Under uncertainty, it is necessary to evaluate all the possible outcomes of all strategies. The basic idea, then, is to identify robust strategies so that, whatever happens, only acceptable world-states are attained. But ensuring robustness is just a version of the maximin-procedure, and so is the tolerable windows approach, too. However, the tolerable windows approach as applied in the TAR seems to assume perfect foresight because only one single climate model is used and because the consequences of each strategy (i.e. the resulting emission-path) are assumed to be known. Disappointingly, the IPCC does not succeed in taking uncertainty fully into account when developing decision making procedures.

Finally, we will discuss the most important practical equivalent of the maximin-rule: the precautionary principle (PP). Morris (2000) distinguishes the strong and the weak PP. As an exemplification of the strong version, he cites the conclusions of the *Wingspread Conference on the Precautionary Principle* from January 26, 1998,

> When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically.

---

6 As we shall see later, this is not the only sense in which the term "robustness" is used in decision theory.

7 Most of the presented approaches require CPFs.

> In this context the proponent of an activity, rather than the public, should bear the burden of proof. (Science & Environmental Health Network, 1998)

This implies that measures to prevent some *possible* harm, if necessary the cessation of the activity that might cause it, must be adopted. Thus, activities are rational only if it is *im*possible that they cause harm to human health or the environment. The strong PP focuses on worst-possible outcomes and is therefore just another version of the maximin principle.

In contrast to the strong PP, its weak version merely asserts that lack of scientific certainty is not a reason to postpone measures which aim at preventing harm. This is the version which predominates in international law; and one of the first formulations numerous treaties now refer to[8] is the 15th principle of the *Rio Declaration on Environment and Development*,

> In order to protect the environment the Precautionary Approach shall be widely applied by states according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation. (United Nations, 1992)

Thus, the weak PP simply asserts what we have taken for granted throughout the last chapters: that we cannot wait until uncertainty is reduced to risk or even certainty before taking a decision. This stems from the fact that we face irreducible uncertainty in many policy domains. We cannot postpone decisions under uncertainty until we obtain credible CFs for the same reasons that we cannot postpone decisions under risk until we obtain credible CFs. A gambler engaging in roulette who tried so would never ever put a stake on any field. Hence, Gollier & Treich (2003, p. 81) are absolutely right to contrast the (weak) PP with a "wait-and-see" strategy. Beck (1986) pinpoints the effects of denying the weak PP,

> By forcing up the scientific standards one minimizes the group of accepted and policy-relevant risks, and as a consequence implicitly issues allowances for risk potentiation. To put it pointedly: Insisting on the purity of scientific analysis leads to the pollution and contamination of air, food, water and soil, plants, animals and men. (p. 86)

Given that the weak PP rather asserts a triviality, it is surprising to see that (i) in the Rio-formulation it applies only conditionally to "threats of serious or irreversible damage" and (ii) it seems to be contested even in its weak form. As to (i), if uncertainty is irreducible, a decision has to be taken based on the available

[8] See for instance European Environmental Agency (2001, pp. 13f.).

knowledge no matter whether the stakes are high or low. And (ii) the objections raised against the weak PP are just gormless: Thus, Morris (2000, pp. 13f.) enumerates five problems of the weak PP — three of which only urge that the terms "threat", "damage" and "serious" have to be defined. Such a critique, besides merely touching the conditions under which the principle shall apply and not the (self-evident) principle as such, is unsustainable for these notions are supposed to be vague and value-laden because their very function is to enable agents to normatively evaluate world-states. Any demand for defining them objectively is completely missing this important point. The fourth problem consists in the fact that every action is strictly spoken "irreversible" — yet this is pedantic because clear cases could still be distinguished even if actions differed *in degree* of irreversibility only. Finally, the fifth problem,

> Fifth, science has not yet, and is unlikely in the future, to provide a fully fledged deterministic theory of the universe from which all particular events can be predicted. In other words, there will always be scientific uncertainty, both with regard to environmental effects and with regard to all other matters, especially concerning the future. (Morris, 2000, p. 14)

Yet that is what we have been urging all the time. And it is the reason why we cannot wait until uncertainties are resolved! The weak PP says no more than that.

We should note that the weak PP (as opposed to the strong one), like the optimism-pessimism principle, is not a bridge-principle and does not connect strategies or decisions to background knowledge under uncertainty. Instead of being deduced from the weak PP, "[the] appropriate response in a given situation is [...] the result of an eminently political decision, a function of the risk level that is 'acceptable' to the society on which the risk is imposed" as the European Commission underlines in a Communication on the PP (European Commission, 2000, p. 15). Nonetheless, such a political decision may of course be based on some bridge-principle such as the maximin principle for instance.

*Problems and limitations*

Whilst the mean approach has faced the problem of insufficient robustness to ignorance, the maximin principle seems to be better off: As generally (unless the lexicographical rule is applied) only one possible outcome per strategy is considered, the rational choice entailed by the maximin principle and some background knowledge will hardly vary if the degree of ignorance changes. Precisely, it will only vary if a new possible outcome of the prescribed strategy arises which is worse than another strategy's worst possible outcome. But although maximin is suited for decision making under ignorance, it faces several other problems as the following five cases will illustrate.

*First case: Gambling.* Imagine you were invited to participate in a TV quiz-show where you might win up to 1 million euro. If you accept, the worst possible outcome is to win nothing, to pay the 500 euro travel cost and, eventually, to be stultified for all the world to see. When, on the contrary, you stay at home and watch the show from your sofa, the worst possible outcome is to be bored. Assuming that being bored for 45 minutes is not as bad as paying 500 euro and being publicly stultified, the maximin principle entails not to participate in the show. However, I am sure that this is not what most of us would consider as the (only) rational choice.

The lesson to be learnt from this case is that focusing on the worst possible outcome only — even when it is quite acceptable — might result in missing worthwhile opportunities. Maximin should therefore be restricted to situations where the worst possible outcome is really intolerable. This is the reason why serious or even irreversible damage is usually considered as a necessary condition for applying the strong PP.

This idea can be put in a more general and more metaphorical way: Maximin assumes that the situation of interest is *fragile* what implies the possibility of bursting, irreparable damage and loss. In such cases, one should try everything to prevent the worst case. Now: How fragile is our economy? How fragile is our global environment? How fragile is life? Answers to these questions determine one's stance towards the maximin principle. If the idea of fragility is central in one's Weltanschauung, if one is aware of the dangers that social or environmental systems might burst, if one perceives the ice on which we are walking to be thin, to put it in one of my teacher's favorite metaphors, then one inclines to apply a precautionary approach.

*Second case: Investment.* Consider an investment decision; the investor typically faces the choice between (i) realizing a project whose future returns are unknown (uncertainty) and (ii) paying his money into a bank account and enjoy constant interest payments. If everything goes fine, (i) results in much higher returns than the interest payments whereas its worst possible outcome usually consists in loosing all one's money. As we lack objective probabilities, maximin can be applied and it entails not to realize the project (i) but to save the money (ii) instead. Given the severely bad possibility of loosing all the money, this choice seems to be reasonable according to our intuitions. However, as all the investment decisions are of this kind and as maximin thus applies generally, the social outcome of the approach is largely undesired: zero investment, economic stagnation and, later, depression, collapse.

*Third case: Scientific research.* When deciding about future scientific research, we face a similar problem. Should we engage in a new scientific research program or not? Although the promises of the research are always high, so are the potential harms, too. The new research might on the one hand give rise to applications which solve the global energy problem, allow food production in the desert, ultimately defeat lethal diseases, as well as, on the other hand, provide methods for mass production of weapons of mass destruction, or trigger off the final ruination of the global environment. With respect to these consequences, we are not only uncertain, but largely ignorant. The alternative to conducting new research is to continue business as usual — admittedly not as bad as those horror scenarios just mentioned. Thus, maximin prescribes not to engage in the new research. Again, whilst this might be intuitively the right thing to do in each single case, certainly not everybody considers its *collective outcome*, this is the deadlock of science and the end of scientific and technological progress, as desirable.

The interesting point about these last two cases is: Even if we take it for granted that maximin can be applied for each single decision (it is rational not to continue this research, it is justifiable not to realize that project and pay the money into an account), the outcome of collectively doing so (scientific deadlock and economic collapse) is undesired! The situation is similar to the prisoners dilemma and the public goods problem where rational behavior on the individual level does not bring about the most preferred situation. In other words, individual rationality does not coincide with what we may loosely call collective rationality. This also seems to be true for the maximin principle: The result of consistently applying maximin is not only different from but sharply opposed to the outcome which would follow if individual behavior were coordinated effectively. Maximin must therefore be applied with care and we cannot urge decision making under uncertainty to be based on a principle with possibly devastating consequences. The problem ultimately arises because the possible outcome of collectively applying maximin (scientific deadlock and economic collapse) does not appear in the individual's $\gamma$-tree as a possible consequence of the option which seems to maximize minimal utility. A solution to this problem must internalize the possible outcome of collective application into the decision making process. Coordination through collective deliberation seems to be one way to do so, but it exceeds the well-structured logico-semantical approach to rational decision making under uncertainty and ignorance which consists in making choices consistent with some background knowledge and a bridge-principle. Here we already catch a glimpse of the direction of thought we will follow in the next chapter.

*Fourth case: Nuclear energy.* In the following discussion about whether to abolish nuclear energy or not, the opponents try to base their policy recommendation on the maximin principle. Neo No argues: "The civil use of nuclear power might sooner or later spark off a maximum credible accident, contaminating the whole continent, causing millions of casualties. We should abolish nuclear power today rather than tomorrow!" "Our technology," replies Peter Pro, "is the safest world-wide. Abolishing nuclear power in this country might lead to a lowering of safety-standards world-wide and thereby to a maximum credible accident. There's no use in abolishing nuclear energy!" And Patty Pro, his sister, adds: "Even if we abolished nuclear power, other countries would continue to use it and the worst case is not banned — if the worst cases are equal for both strategies, why then opting for one rather than for the other?" Neo No is just about to reply when George H. Green (member of Environmentalists For Nuclear Energy) intervenes: "Look, abolishing nuclear power world-wide — if that is what you wanted to suggest — might lead to such an increase in GHG-emissions that, through global warming, the ice caps melt and sea levels rise. Moreover, melting hydrates trigger off submarine landslides: A tsunami of unseen size destroys coastal areas and cities comprehensively all over northern Europe. In addition, the Gulf Stream will collapse because the release of large amounts of fresh water interrupts the thermohaline circulation. As a consequence, not only the already devastated coastal regions but the whole of Western Europe will suffer from drastic regional climate changes. As to the other continents ..."

This discussion illustrates that there are situations where for each option there is a consequence which is as bad as the other strategies' worst possible outcomes. The nuclear energy case is certainly not unique in this sense: As soon as the decision problem is sufficiently complex, we can imagine worst-case scenarios abundantly.[9] This, however, implies that maximin does not fulfill its guiding function in sufficiently complex situations — a serious limitation to our principle. Our example shows that the $\gamma$-tree can be so large, every optional strategy including so many different possible outcomes, that there is no difference between strategies with respect to their worst possible outcome. In fact, if maximin cannot be applied, strategies also have to have similar second worst possible outcomes, third worst possible outcomes, etc. because a lexicographical ranking could be established otherwise. But then, there is no difference between the options as represented by the $\gamma$-tree at all and no rule can differentiate between them. In such

[9] This is also noted by Hansson (1997) in the course of his discussion of the PP.

cases, there seems to be no distinction between rational and arbitrary decision making.

*Fifth case: The tube.* We are at Green Park, London and want to take the tube to King's Cross/St. Pancras: either Victoria Line (4 stations) or Piccadilly Line (6 stations). However, as we hurried to the platform, we have no idea whether the Victoria Line is still interrupted between Euston and King's Cross. Thus, taking the Victoria might result in not reaching King's Cross in time. Maximin hence prescribes to take Piccadilly Line at Green Park. But there is no need to take the slower Piccadilly Line because even if Victoria Line were interrupted, we could change at Euston to the Northern Line to reach our destination.

The maximin as we have discussed it so far does not explicitly consider the possibility that initial decisions might be corrected later on so that the worst possible case can be mitigated. This, however, is not a shortcoming of maximin but of our framework which focused on single decisions rather than on decision sequences. We will now introduce the framework of sequential decision making and show how maximin can be applied therein.

## 12.4 Decision sequences and maximin

The decision situations we have dealt with up to now were of this kind: An agent has to take one single decision involving alternative strategies which have several possible outcomes. We have represented such situations by the — as we should call it from now on: — *normal* $\gamma$-tree (figure 11.2). Now, many real problems we face do not merely require one single decision before they can be filed away. Policy-making is frequently analyzed as a policy-cycle involving continuous decision making. The importance of conceiving decision making as a continuous process has also been highlighted in climate policy,

> Such a sequential decision-making process aims to identify short-term strategies in the face of long-term uncertainties. The next several decades will offer many opportunities for learning and mid-course corrections. The relevant question is not "what is the best course of action for the next 100 years", but rather "what is the best course for the near-term given the long-term objective?" (IPCC, 2001*a*, p. 613)

The point, however, is not that such a situation does not fit into the normal $\gamma$-tree. Any decision does. The point is rather that the normal $\gamma$-tree only contains further decisions implicitly in the possible world-states instead of making them explicit. The structure of the normal $\gamma$-tree can yet be expanded so that all (relevant) future

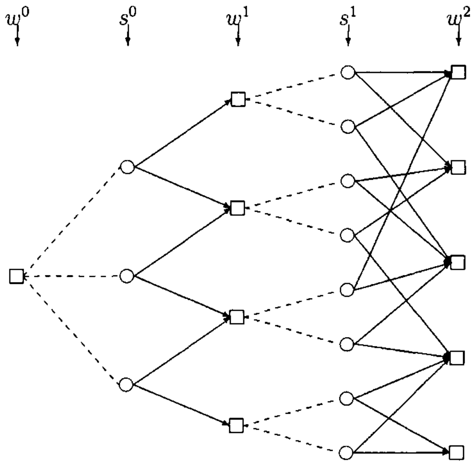



Figure 12.1: An expanded $\gamma$-tree of depth 2, representing the uncertain outcomes of decision sequences.

decisions are explicit. Each possible outcome of some strategy becomes the origin of a further decision problem. The result of this procedure is the *expanded* $\gamma$-tree as shown in figure 12.1.

The expanded $\gamma$-tree is divided into several layers. The very left layer, the $w^0$-layer, contains the original decision point, i.e. the root of the whole tree. The alternative strategies to be chosen make up the next layer, the $s^0$-layer. Their possible outcomes are displayed in the $w^1$-layer and so on. For practical purposes, the expanded $\gamma$-tree has to be finite. Its *depth* is the index-number of its last $w$-layer. We say that this layer contains the *final states* of the decision problem. Preferences or utilities in sequential decision problems are given for these final states only.

Now, which strategy should be adopted today ($w^0$-layer) in the light of future decisions and possible outcomes according to the maximin rule? As we do not know the utilities of the possible outcomes, $w_i^1$, of the alternative strategies at layer $s^0$, we cannot apply maximin directly. Nevertheless, we can determine a choice recursively by evaluating the strategies (world-states) at layer $l$ given the evaluation of the world-states (strategies) at layer $l+1$, starting at the final states. Maximin may then serve as one of the recursive rules which define an evaluation function $v()$ (for *v*alue) over the expanded $\gamma$-tree,

R1 Initial layer: $v(w_i^n) = U(w_i^n)$ for all final states $w_i^n$.

R2 Evaluate $s$-layer: $v(s_i^l) = \min_{j_i}(v(w_{j_i}^{l+1}))$ for all strategies $s_i^l$ in layer $s^l$ and their corresponding possible outcomes $w_{j_i}^{l+1}$.

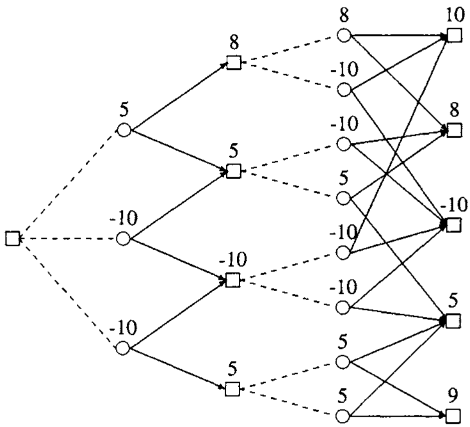



Figure 12.2: Evaluating an expanded $\gamma$-tree with recursive maximin rules.

R3 Evaluate $w$-layer: $v(w_i^l) = \max_{j_i}(v(s_{j_i}^{l+1}))$ for all world-states $w_i^l$ in layer $w^l$ and the resulting alternative future strategies $s_{j_i}^{l+1}$.

The first rule simply sets the value of the final states to the corresponding utilities we have assumed to be given. The real recursive rules are R2 and R3. R2 declares how to evaluate a strategy given the evaluation of its possible outcomes. According to R2, a strategy is worth the value of its worst possible outcome. This is the maximin rule. To complete the recursive definition of $v()$, R3 defines how to evaluate a world-state $w$ given the evaluation of the alternative options available under $w$. Here, the decision maker is assumed to opt for the best available strategy given some alternatives so that the world-state is worth the value of the best strategy it makes available. With these rules, the values of the strategies in the $s^0$-layer can be determined. The rational agent then adopts the strategy with the highest value.

Figure 12.2 exemplifies the evaluation of an expanded $\gamma$-tree with the recursive rules just introduced. A handy way to characterize the maximin rule under sequential decision making is: Assume that there is a malicious daemon who influences the possible outcomes to your detriment. This is of course just a convenient way of talking — adopting the maximin rule does not entail (having the belief) that there is such a daemon. Also note that the crucial recursive rule is R2. In fact any bridge-principle originally designed for the normal $\gamma$-tree we have discussed so far can be applied to the expanded $\gamma$-tree by plugging it in as the recursive rule R2.

Applying maximin to sequential decision problems has helped us to get rid of the problem illustrated by the fifth case above. The other problems, however, still remain. The easiest way to see this is by applying the recursive maximin rule to expanded $\gamma$-trees of depth one. In this case, single and sequential decision making are identical and so are the original and the extended maximin rule. Hence, whatever poses a problem for the normal maximin approach, does so for recursive maximin, too.

Before we leave the maximin approach and investigate another family of bridge-principles, we should consider the scientific requirements if decision makers adopt the maximin-rule. This is fairly obvious: Under maximin, the most important parts of our background knowledge are those which depict worst case scenarios. Scientists should thus give priority to the investigation and identification of very bad, intolerable possible outcomes of available strategies.

## 12.5 Counting possibilities and future options

The bridge-principles which we will discuss in this section connect the decision in favor of a strategy to the background information about how many different possible developments this strategy gives rise to and how many future options will be available once the strategy is implemented.

*Minimizing future possibilities*

This bridge-principle prescribes to choose a strategy so as to minimize the degree of future uncertainty,

**Bridge-principle 6 (Minimize uncertainty)** *The rational agent adopts the strategy with a minimum of possible future outcomes.*

This principle intuitively makes sense because the less possible outcomes, the better we can prepare for each of them. Consider as an example a transport problem: You are charged with transporting three items with a lorry to a remote village. All three items, namely a barrel of gasoline, a barrel of oil and a container full of mechanical tools, are urgently needed and not receiving one would cause equal damage to the local population. Now, there are two distinct routes you can take, both lead you through areas which are controlled by gangs. Route A112 takes you through the swampland and you might be attacked by a gang which is always keen on oil and oil only. Taking the other route, H64, you might be ambushed in the rainforest but you know nothing about the rainforest gang's preferred booties. Yet what we know for sure is that whatever route you take, there will be no more than one raid and if there were a raid indeed, the gang would not steal more than one item. So, what are you going to do? A sensible choice seems to be: Call the local villagers, inform them that the oil might not arrive as planned, and than take route A112 through the swampland. This seems to be sensible because under this choice the villagers at least have the chance to prepare for the missing oil — whereas if you took route H64, they would have to prepare for a possible lack of any of the items. This is a sound application of the minimize uncertainty principle.

In order to apply minimize uncertainty to sequential decision problems, we have to define a set of recursive rules. As we evaluate strategies according to the set of future possibilities they give rise to, we shall adopt a set-theoretical approach and define the evaluation function $v()$ over the expanded $\gamma$-tree as follows,

R1 Initial layer: $v(w_i^n) = \{w_i^n\}$ for all final states $w_i^n$.

R2 Evaluate $s$-layer: $v(s_i^l) = \bigcup_{j_i}(v(w_{j_i}^{l+1}))$ for all strategies $s_i^l$ in layer $s^l$ and their corresponding possible outcomes $w_{j_i}^{l+1}$.

R3 Evaluate $w$-layer: $v(w_i^l) = \min^*_{j_i}(v(s_{j_i}^{l+1}))$ for all world-states $w_i^l$ in layer $w^l$ and the resulting alternative future strategies $s_{j_i}^{l+1}$.[10]

Each final state gives rise to one possibility — the final state itself (R1). A strategy gives rise to as many possible final states as all its direct possible outcomes together (R2). And, as previously, we assume that the decision maker chooses strategies so as to minimize future possibilities (R3). The evaluation function $v()$ attributes the set of future possibilities to each node in the expanded $\gamma$-tree and the minimize uncertainty principle therefore prescribes to chose that strategy at $s^0$ whose value has the smallest cardinality. For illustration, the recursive rules are applied in figure 12.3.

The limitations of the minimize uncertainty approach, however, are numerous. First of all, it contradicts the maximin rule in many cases where the latter can be applied sensibly. Even more, minimize uncertainty completely disregards preferences and utilities. This indicates that it should be restricted to situations where considering utilities is not helpful or has become illusive, cases as exemplified by the transport problem (equal utilities of possible outcomes) or the nuclear energy discussion ($\gamma$-tree too complex). Furthermore, minimizing uncertainty is not quite as robust to ignorance as the maximin approach because the minimize uncertainty approach considers the cardinalities of all alternative sets of possible outcomes and these cardinalities are always altered by a reduction or an increase in ignorance.

These limitations appear to be negligible in the light of the most severe flaw of the minimize uncertainty principle. Maybe the seemingly sound applications of this approach are no application of it at all. Consider the following example to see

[10] Here, $\min^*_i(a_i) := a_k$ with $|a_k| = \min_i(|a_i|)$.

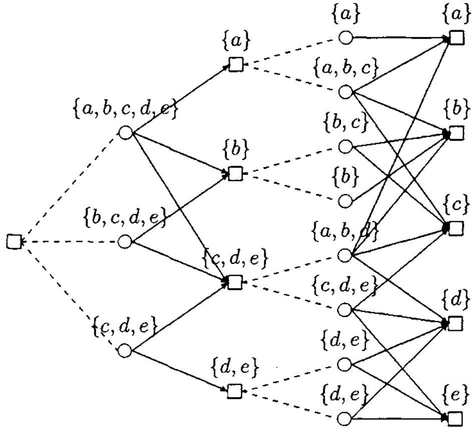



Figure 12.3: Evaluating an expanded $\gamma$-tree accordingly to the minimize uncertainty principle.

why the possibility of preparation is a core condition of the minimize uncertainty approach: You are given four chips: blue, red, yellow, green. If you choose option A among two alternatives (A and B), you possibly loose the green chip. Choose B and you possibly loose the green, the yellow *or* the red chip. For each chip you possess at the end of this game, you get 100 euro. Now, which option do you chose? A, because there are less different possible outcomes? Yet nothing seems to be wrong with choosing B instead. But if choosing B is as reasonable as opting for A, then something has to be wrong with the bridge-principle! In contrast to the transport case, there is no way one could prepare for one of the possible outcomes of the game. Via the fixed rewards, equal utilities are guaranteed. On the other hand, preparation actually means to change the utility of an outcome. We call the villagers so that they can start saving oil and thereby reduce the damage of not receiving the supplies. If that is the correct analysis, the worst possible outcome of taking route A112 (oil is stolen, but villagers are prepared) has a higher utility than the worst possible outcome of taking the H64 (an item is stolen but villagers are not prepared) and the intuitively sensible choice for A112 rather results from an application of the maximin principle than from the minimize uncertainty approach.

Yet, in cases where preparation is feasible and might reduce the harm of some possible consequences, but where this harm-prevention is for some reason not included in the world-states' description and therefore is not reflected in the utility function, the minimize uncertainty approach might serve as a useful and simple *heuristic* to incorporate the preparation-reasoning. As such, it should not replace other bridge-principles, but rather complement them where they do not lead to unique prescriptions.

*Maximizing future options*

An approach to decision making under uncertainty we are all familiar with is to try to 'keep one's options open until one knows more'. Such an approach which favors flexibility and openness makes sense under the crucial assumption that uncertainty will be reduced in the future. If by tomorrow we will know as little as today, then there is no point in postponing a decision. If, on the other hand, uncertainties are going to be reduced significantly (in degree), we should try to delay the decision in order to explore the new information in the future. Notice that what counts is not the absolute degree of uncertainty but its relative decrease.

Just-in-time decision making is the perfect realization of the openness approach: If flexibility is so high that decisions can be taken without any planning- and thus forecast-horizon, then uncertainty is no longer an obstacle to rational decision making because the standard approach (bridge-principle 1) can be applied to the just-in-time decision. But just-in-time decision making requires that uncertainties will be entirely reduced and that by that time all alternative outcomes are still 'reachable'. Although these conditions will hardly be satisfied in the policy domains we primarily deal with, they depict the state the openness approach aims at.

Since the openness approach urges to make today's decision in the light of future decision options, it applies to sequential decision problems only. Figure 12.4 illustrates the important assumption of the openness approach, namely that some of the arrows connecting the strategies in the last $s$-layer with the final states will disappear once the initial decision is made and time goes by.

Now, how do we state the openness approach in terms of recursive rules that allow us to evaluate an expanded $\gamma$-tree? At first glance, the openness principle seems to state that we should try to maximize future options, i.e. the number of available strategies, and that a world-state is worth the number of strategies it makes available. In spite of being the underlying idea this is too simple: First of all, not every further strategy makes a world-state more valuable. Some strategies might for instance inevitably cause a worst-case and should not be considered as increasing the value of a world-state. Secondly, the proposed rule would not result in a recursive evaluation of the expanded $\gamma$-tree. Each $w$-layer could be evaluated independently of the rest of the tree. Yet there is a straight forward way to fix these problems: The value of a world-state is defined as the sum of the values of all directly available strategies. This yields the recursive rule R3.

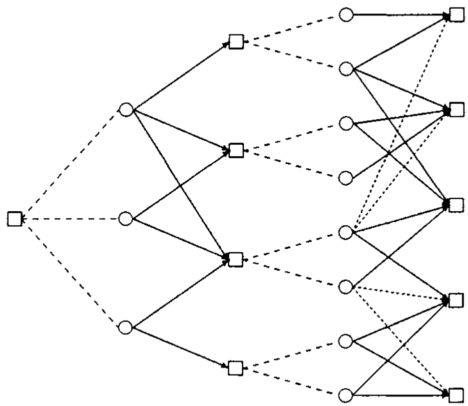


Figure 12.4: The assumption that uncertainty will be reduced illustrated in an expanded $\gamma$-tree: Some of the arrows to the final layer will eventually disappear.

What about the missing rules R2 and R1? R1, of course, is unproblematic and can be defined as before. R2, however, poses a problem. The question is: What makes a future strategy valuable under the assumption that some of its possible outcomes might turn out to be impossible? I guess any of these criteria could be considered: the number of intolerable possible outcomes, the number of highly preferred possible consequences, the number of possible outcomes, the utility of the best possible outcome, and so on. Hence, the openness approach does not provide us with a clear instruction of how to evaluate strategies given the values of their possible outcomes. But defining R2 has always been the most difficult task when setting up a set of recursive rules and that was what we have used the bridge-principles for. Now if the openness principle does not determine R2, we may simply use any other bridge-principle to do so. Thus, the openness approach can be combined with maximin, optimism-pessimism, minimize uncertainty or any combination of them.

In the following set of recursive rules, the maximin rule completes the openness approach,

R1 Initial layer: $v(w_i^n) = U(w_i^n)$ for all final states $w_i^n$.

R2 Evaluate $s$-layer: $v(s_i^l) = \min_{j_i}(v(w_{j_i}^{l+1}))$ for all strategies $s_i^l$ in layer $s^l$ and their corresponding possible outcomes $w_{j_i}^{l+1}$.

R3 Evaluate $w$-layer: $v(w_i^l) = \sum_{j_i} v(s_{j_i}^{l+1})$ for all world-states $w_i^l$ in layer $w^l$ and the resulting alternative future strategies $s_{j_i}^{l+1}$.

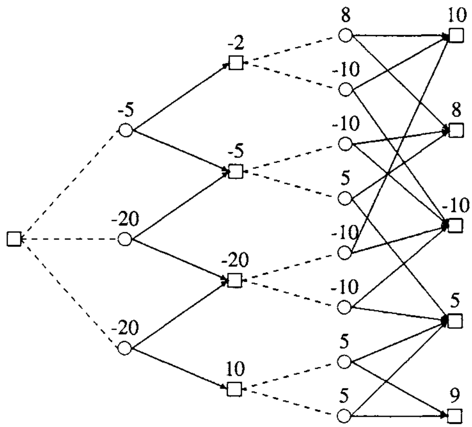



Figure 12.5: The openness approach and the maximin principle are combined to evaluate the expanded $\gamma$-tree.

Figure 12.5 illustrates these rules by applying them to a fictitious decision problem.

The fact that the openness principle requires a further bridge-principle and can be combined with maximin is maybe less surprising if we consider Beck's argumentation in favor of the openness approach,

> Alternative developments can thus obstruct the future or leave it open. One either takes a decision for or against a journey to the unknown no man's land of the though unseen but foreseeable 'side-effects'. Once the train has departed it is difficult to stop it. We hence have to choose alternative developments that do not obstruct the future and that transform the process of modernization itself into a process of learning by ensuring that the withdrawal of lately discovered side-effects is always possible through the reversability of decisions. (Beck, 1986, p. 294)

Now Beck suggests that this openness is ultimately linked to a worst-case approach,

> We thus have to test practical developments for a risk giantism which would deprive men of their humanity and condemn them to faultlessness from now on to all eternity. (p. 293)

The argument asserts that the worst case entails that there is no way back and that the remaining options are worthless. The worst case is identical with the no-more options case. Trying to prevent worst-cases (following maximin) is therefore, according to Beck, necessary for safeguarding and maximizing valuable future options.

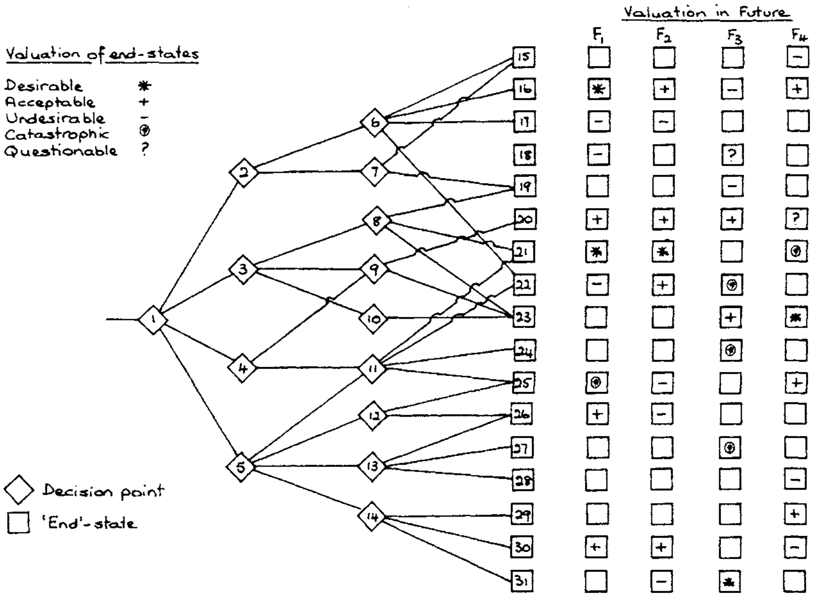



Figure 12.6: Rosenhead's decision tree. (Source: Rosenhead, 2001*a*, p. 196)

*The other side of the coin: Rosenhead's robustness analysis*

We have noted in the previous chapter that uncertainty about a strategy's outcomes can always be rendered as uncertainty about the evaluation of a (certain) outcome. Here comes a direct application of this observation: If we transform the expanded $\gamma$-tree so that the evaluation of the final states becomes uncertain, the openness approach presented above is almost equivalent to the "robustness analysis" of Rosenhead (2001*a*). I will first present Rosenhead's analysis and then show how exactly it relates to our openness approach. Henceforth common limitations of each method will be discussed.

I shall briefly introduce Rosenhead's approach alongside the diagram he uses himself for illustration (figure 12.6). The decision tree is composed of different layers, each consisting of "decision points" except the last one which comprises the "end-states". Decision points are best interpreted as world-states which the decision maker can bring about. The connections between the decision points indicate which world-states can be directly attained from some decision point. World-state 8, for instance, can directly be brought about from world-state 3 whereas world-state 7 cannot. Thus, the decision maker is assumably completely controlling and determining which world-state is reached. In contrast, uncertainty governs the evaluation of the end-states. Rosenhead considers finitely many different

| Initial decision | Future $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|---|---|---|---|---|
| 2 | 1/5 | 2/5 | 0 | 1/4 |
| 3 | 2/5 | 2/5 | 2/3 | 1/4 |
| 4 | 2/5 | 3/5 | 2/3 | 2/4 |
| 5 | 3/5 | 3/5 | 1/3 | 2/4 |

Table 12.2: Robustness scores of initial decisions under alternative future scenarios

"futures" which give rise to different evaluations of the end-states as represented by the four columns on the very right of figure 12.6. Hence, he implicitly assumes that future developments are external to the decision making process.

Rosenhead defines "robustness" of an initial decision ("commitment") as "the ratio of the number of acceptable configurations that are reachable from that commitment, to the total number of configurations which have been identified as having acceptable expected performance at the planning horizon" (p. 189). Thus, "the robustness of each candidate initial decision package can be calculated for each future separately" (p. 191). The robustness score of each initial decision according to the decision tree in figure 12.6 is given in table 12.2. This is the information the decision has to be based on. How? Rosenhead's robustness analysis does not tell us. In fact, table 12.2 can be represented by a normal $\gamma$-tree, and we still face a classical one-step decision problem under uncertainty of the kind we have been dealing with throughout this and the previous chapter. And we therefore need a further bridge-principle which guides our decision. This can be the maximin rule or any other principle. Just like the openness approach, Rosenhead's robustness analysis has to be completed by a further bridge-principle, and is compatible with any of them.

This said, how can we make sense of Rosenhead's approach in terms of our framework? As mentioned above, I suggest to interpret decision points as world-states which can be brought about by the decision maker. Thus for each of these world-states a strategy is available which surely gives rise to it. As described in the previous chapter, the uncertainty about the evaluation of end-states can be interpreted as an uncertainty about the further consequences of the end-states. In contrast to Rosenhead's analysis, this uncertainty is necessarily internal to the decision making process in our framework, i.e. there is no a possible future for which all end-states can be evaluated but there are, for each end-state, different possible futures. Figure 12.7 is then a representation of a typical decision problem à la Rosenhead by an expanded $\gamma$-tree.

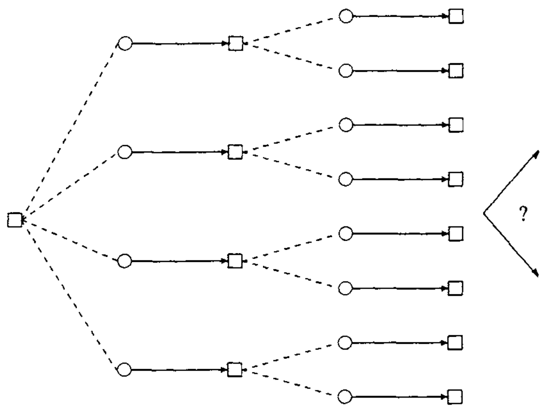



Figure 12.7: A typical decision problem Rosenhead considers represented by an expanded $\gamma$-tree. The uncertain consequences of the final-states induce the different possible evaluations.

*Problems and limitations*

The openness approach is without doubt a valuable approach to decision making under uncertainty. As an example, we have already mentioned just-in-time approaches which are for instance proposed for management decisions by Bryan (2002) and Beinhocker & Kaplan (2002). Rosenhead (2001*b*) presents another sound application where the decision problem consists in choosing the subjects for an O-level examination. In spite of these sound applications, the openness approach faces some problems and limitations.

The first limitation arises out of the central assumption of the openness approach which might indeed be problematic: that uncertainty will be reduced by degree in the future. Consider for example climate change. It is not clear at all that we will have substantially reduced our uncertainty about the long term impacts of GHG-emissions on our climate in the relevant future, say the next 20 years. Alas, postponing the decision might not only be useless but harmful because we might have lost 20 important years to prevent severe consequences of climate change. Thus, before applying the openness approach it is crucial to check whether uncertainties may be reduced or not. This, of course, is itself uncertain.

We have seen that both versions of the openness approach require an additional bridge-principle for completion. This bridge-principle had to specify the second recursive rule, or the final choice given the robustness-scores (table 12.2) respectively. But we have not yet identified an absolutely unproblematic bridge-principle and every limitation or problem of the bridge-principle used to complete the openness approach becomes a limitation or problem of that very approach, too. If we use for example the maximin principle, then all its problems limit the

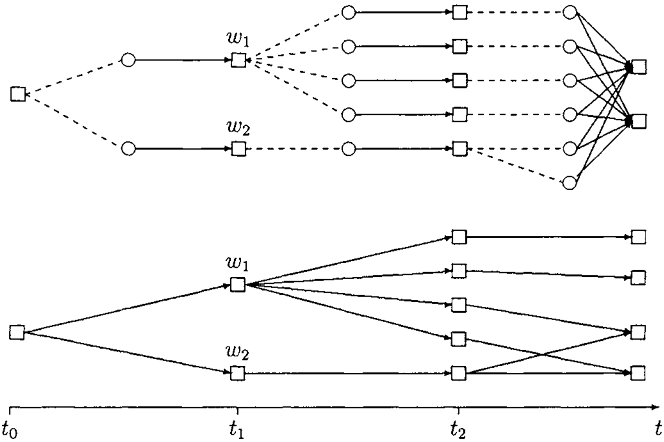



Figure 12.8: Two decision trees illustrating time-inconsistencies of Rosenhead's robustness analysis (bottom) and the openness approach in our framework (top).

applicability of the openness approach.

Furthermore, the openness approach faces so-called time inconsistencies. Consider the two decision trees in figure 12.8. No matter what the evaluations of the end states and final states, the openness approach complemented by maximin favors the decision to attain world-state $w_1$ rather than $w_2$.[11] And indeed, opting for $w_1$ seems to make sense assuming that uncertainties will be at least partly resolved at time $t_1$. But what if uncertainties are not resolved by that time and will be reduced later, $t_r > t_1$? Then, having chosen $w_1$, we have to reduce our future options at the second decision step ($t_1$) to only one without knowing more than at $t_0$. If we had chosen $w_2$ instead of $w_1$, we would guard at least two options for $t_2$ and thus, in terms of flexibility, be better off! If time $t_r$ (when uncertainty is reduced) is greater than $t_1$ the openness approach might ultimately lead to a non-optimal prescription.

This problem, however, may be fixed by modifying the approaches slightly. Rosenhead's approach can be modified so as to evaluate the robustness-score of the decision points at time $t_r$ rather than of the initial decisions. The agent then makes an initial decision which allows him to reach the decision point with the highest robustness score at $t_r$. Regarding the evaluation of the expanded $\gamma$-tree, the recursive openness rules have to be restricted to the evaluation of the layers at $t \geq t_r$. The remaining layers are then evaluated according to the recursive rules of the sequential maximin approach. Both modifications will make sure

[11] It is easy to verify that the robustness-score of $w_1$ is greater than or equal to that of $w_2$ and that $v(w_1) = 2v(w_2)$.

that, if $t_r = t_2$, the initial decision in the example will be $w_2$. We can finally note that this modification underlines the importance of the uncertainty-reduction assumption. If this assumption is not satisfied, i.e. $t_r \rightarrow \infty$, then the modified openness approach becomes nothing but the sequential maximin procedure.

As to the robustness to ignorance, there are two points to be considered. Like for the other bridge-principles, we have to figure out whether the choice entailed by the openness approach alters when ignorance is introduced. Besides this, there is the additional question in how far the crucial assumption of the openness approach is undermined by ignorance. Let us start with this one. At first glance, the introduction of ignorance into the decision problem seems to spoil the assumption that uncertainty will be reduced. For under ignorance, uncertainty will possibly grow as new, unseen possibilities are discovered. On the other hand, maximizing valuable future options (available strategies) might even be a sound strategy under the assumption that some currently known possibilities might turn out to be impossible (reduction of uncertainty) *and* some yet unknown new possibilities might emerge (reduction of ignorance). For if one has acted so as to maximize the amount of valuable options and an unforeseen intolerable possibility arises, there might still be some valuable options which exclude that worst case. If, in contrast, the newly emerged possibility is highly attractive, some of the safeguarded options might include it as a possible outcome. This said, ignorance does not undermine the assumptions of the openness approach, but rather strengthens its case.

Yet the question remains whether the choice prescribed by the openness approach alters given a reduction of ignorance. This primarily depends on which bridge-principle is used to complete the approach. Accordingly, a combined maximin-openness approach is much more robust to ignorance than a combination of the mean principle and the openness approach. All things considered, the combination of maximin and openness seems to be sufficiently robust to ignorance.

## 12.6 The organism metaphor

This section discusses attempts to obtain bridge-principles for decision making under uncertainty and ignorance by reasoning by analogy. The volume on tools for policy analysis of the collection *Human Choice and Climate Change* edited by Rayner & Malone (1998) includes a whole chapter on reasoning by analogy (Meyer et al., 1998). Its analogies are drawn between current climate change on the one hand and (i) past climate change or (ii) urban climate change on the other hand in order to learn from how agents dealt with and adapted to the resulting challenges. Thus, analogies are used to directly infer prescriptions about what to do given the uncertain future of our climate. The bridge-principle involved is something like: Behave in analogy to agents in relevant, sufficiently similar cases, if the agents had been successful. Such a principle is, as I see it, only of very limited use as 'sufficiently similar cases' might not be available and the agent's success might simply be due to luck. So instead of using analogies to derive specific prescriptions directly, we will attempt to obtain a general bridge-principle by reasoning by analogy.

But an analogy to what? We need a class of decisions (i) which we frequently face, (ii) which are typically made under uncertainty, and (iii) which we finally manage quite successfully. There is a type of familiar decision situations which satisfy these conditions: We face uncertainty every day when we try to get along with plants, animals and our fellow men, and we are generally quite successful in dealing with them. So, the idea is to observe which principles regulate our acquaintance and contact with organisms in order to see in how far these principles can be applied to other decision contexts, namely to climate and economic policy-making.

*The gardener paradigm*

Though each of us to some extent deals with organisms of different sorts every day, some people have specialized in doing so. For example: the gardener and the doctor. Both have to take decisions under uncertainty; for instance when they diagnose unfamiliar symptoms and have to decide how to treat the plant or the patient.[12] Their approaches to decision making are certainly distinct from those of an engineer and can thus be characterized negatively: Gardeners and doctors do generally not carry out narrowly focused intervention into the system in order to reach a specific aim. "Well, let's fix the X-component in order to increase the pressure of the Y-tank, then everything will be fine again", is for instance something we are used to hear from engineers, not from a gardener or a good doctor. Here is an attempt to enumerate some positive characteristics of what we shall call the gardener paradigm to decision making under uncertainty.

*Self regulation.* Gardeners and doctors are aware of the system's (as the case may be: limited) capacity to reach the desired state on its own, e.g. to cure itself.

*Vital processes.* Gardeners and doctors support those vital processes of the system which are crucial for self regulation and for the attainment of the desired system state.

*Soft methods.* They preferably do so by creating a suitable environment instead of intervening directly in the system.

[12] On the other hand, not all decisions gardeners or doctors take are decisions under uncertainty, treatments of well known diseases rather involve decisions under risk.

*Parallelism.* Thereby, they implement different measures simultaneously and do not rely on one single measure only.

*Long-term perspective.* Gardeners and doctors consequently adopt a long-term perspective when trying to reach their aims: The system requires constant care instead of a single repair.

The bridge-principle they follow is simply,

**Bridge-principle 7 (Gardener paradigm)** *The rational agent chooses the strategy which satisfies best the (five) characteristics of the gardener paradigm.*

*Gaia theory*

The gardener paradigm is famously applied to the global ecologic system in Gaia theory. Gaia is the shorthand for the hypothesis "that the biosphere is a self-regulating entity with the capacity to keep our planet healthy by controlling the chemical and physical environment" as the architect of this theory James Lovelock (Lovelock, 1987, p. xii) defines it.

Lovelock identifies three principles which are crucial for living with Gaia which partially reformulate, partially extend the above characterization of the gardener paradigm,

1. The most important property of Gaia is the tendency to keep constant conditions for terrestrial life. Provided that we have not seriously interfered with her state of homoeostasis, this tendency should be as predominant now as it was before man's arrival on the scene.
2. Gaia has vital organs at the core, as well as expandable or redundant ones mainly on the periphery. What we do to our planet may depend greatly on where we do it.
3. Gaian responses to changes for the worst must obey the rules of cybernetics, where the time constant and the loop gain are important factors. Thus the regulation of oxygen has a time constant measured in thousands of years. Such slow processes give the least warning of undesirable trends. By the time it is realized that all is not well and action is taken, inertial drag will bring things to a worse state before an equally slow improvement can set in. (Lovelock, 1987, p. 127)

Acting in accordance with these principles implies to minimize interferences with Gaia as far as possible for in the absence of severe interventions conditions suitable for life will be maintained automatically (homoeostasis), whereas results of interventions might be irreversible (principle 2). It follows in particular not to harm the vital subsystems of the Earth, as for instance the rainforests (principle 3).

*Gardening the economy*

This organism metaphor is as familiar in economics as in ecology. Different economists have conceived the economic system as an organism rather than as a huge "socioeconomic machine" from the very beginning of economic science. Consider as a first example Alfred Marshall who used the organism metaphor extensively and in different varieties throughout his *Principles of Economics* (Marshall, 1920),

> [Economists], like all other students of social science, are concerned with individuals chiefly as members of the *social organism*. (I.II.28)
>
> ...But the chief outcome of recent studies is to make us recognize more fully, than could be done by any previous generation, how little we know of the causes by which progress is being fashioned, and how little we can forecast the ultimate destiny of the *industrial organism*. (I.IV.27)
>
> ...The main drift of this study of Distribution then suggests that the social and economic forces already at work are changing the distribution of wealth for the better: that they are persistent and increasing in strength; and that their influence is for the greater part cumulative; that the *socio-economic organism* is more delicate and complex than at first sight appears; and that large ill-considered changes might result in grave disaster. (VI.XIII.53) [all italics added]

The organism metaphor is not only used as a mere illustration but also to generate new economic insights by reasoning by analogy, and Marshall even applies Darwin's theory to explain the evolution of the "industrial organism" (IV.VIII.2-6).

Yet already the founding father of modern economics, Adam Smith, made explicit use of the organism metaphor in his *Wealth of Nations* and even applied the idea of homoeostasis, i.e. the capacity of self-regulation, to the economic system,

> Some speculative physicians seem to have imagined that the health of the human body could be preserved only by a certain precise regimen of diet and exercise, of which every, the smallest, violation necessarily occasioned some degree of disease or disorder proportioned to the degree of the violation. Experience, however, would seem to show that

> the human body frequently preserves, to all appearances at least, the most perfect state of health under a vast variety of different regimens; even under some which are generally believed to be very far from being perfectly wholesome. But the healthful state of the human body, it would seem, contains in itself some unknown principle of preservation, capable either of preventing or of correcting, in many respects, the bad effects even of a very faulty regimen. Mr. Quesnai, who was himself a physician, and a very speculative physician, seems to have entertained a notion of the same kind concerning the political body, and to have imagined that it would thrive and prosper only under a certain precise regimen, the exact regimen of perfect liberty and perfect justice. He seems not to have considered that, in the political body, the natural effort which every man is continually making to better his own condition is a principle of preservation capable of preventing and correcting, in many respects, the bad effects of a political oeconomy, in some degree, both partial and oppressive. Such a political oeconomy, though it no doubt retards more or less, is not always capable of stopping altogether the natural progress of a nation towards wealth and prosperity, and still less of making it go backwards. If a nation could not prosper without the enjoyment of perfect liberty and perfect justice, there is not in the world a nation which could ever have prospered. In the political body, however, the wisdom of nature has fortunately made ample provision for remedying many of the bad effects of the folly and injustice of man, in the same manner as it has done in the natural body for remedying those of his sloth and intemperance. (Smith, 1776, IV.9.28)

Nowadays, as far as I can judge, the organism metaphor has largely receded and been replaced by the engineer paradigm. The term "Gemeinschafts*diagnose*" (collective diagnosis) which refers to the collective growth forecast of the leading National Institutes in Germany alludes to the medical analogy and may count as a terminological relict of the gardener paradigm.

Once the gardener paradigm is adopted, the implications for economic policy-making are straightforward. To put it negatively first, we cannot and should not attempt to "engineer" economic recovery[13] or any other desired macroeconomic state as we engineer steam engines with some desired properties. As to the positive general prescriptions, we should for instance trust in the self-regulating capacities of the economy and limit economic policy interventions to a minimum. Those readers who have become suspicious when I cited Smith will now certainly cry out that this is a piece of ultra-liberal advice and that the gardener paradigm

[13] As is frequently suggested and expected, for instance by Crooks & Major (2003).

promotes an unsocial and irresponsible laissez-faire! Still, they err. The gardener paradigm in contrast acknowledges the importance of politics: a garden without a gardener is not a garden, but a mess. Similarly, there can be no functioning market economy without any political regulation. According to the organism metaphor, economic policy should consist in setting the framework and creating a suitable environment so that the economy evolves into the desired direction rather than directly intervening in the system. Perturbations can be minimized if measures are made public long before they are implemented and policies are adopted according to public rules rather than discrete decisions. From this point of view, government, in taking care of the economy, is crucially important: In this light the gardener metaphor is anything but ultra-liberal.

But is the gardener metaphor possibly conservative in the sense that it favors the status quo? In fact, it isn't, either. Organisms evolve, can grow, and change their shape (the caterpillar becoming a butterfly). The gardener metaphor simply urges that such changes shall occur gradually as the system evolves at its own pace, and shall not be forced by intervention. Clearly, the evolution can be influenced by creating and changing the system's environment. The gardener paradigm is not conservative, but favors stepwise reforms over revolutions. Regarding the possible evolutions of the economic system, there is something we should not forget: the economic system, in contrast to plants or human bodies, is entirely man-made. Not only in the sense that we have created its rules and regulations, but also in the sense that individual decisions fully determine macroeconomic processes at every instance in time. This implies an enormous plasticity: We can set the rules of the game and individual behavior can change, too. The socioeconomic organism has much more different potential evolution-paths than a biological organism: There is an unlimited number of possible ways how we can organize our social and economic lives. Altogether, system change is possible under the gardener paradigm.

Even if the gardener paradigm is adopted, we have to admit that it is, like all the other approaches discussed, of limited use only. Whilst the organism metaphor only says which *type* of strategy to apply, it does not prescribe a single strategy. According to the gardener paradigm, we should try to bring about the desired world-state by non-interventionist methods — but which of the non-interventionist strategies to choose is still an open question. Thence many different strategies are generally compatible with the given background knowledge and the gardener paradigm. Lovelock (1987) puts this, in the very end of his book, into a pointed conclusion:

> There can be no prescription, no set of rules, for living with Gaia. For each of our different actions there are only consequences. (p. 140)

And the very same conclusion applies *mutatis mutandis* to the socioeconomic organism.

## Final remarks on this chapter

Throughout this chapter we have encountered a large variety of different approaches to decision making under uncertainty. Some of them have been rejected entirely. Others could be reasonably applied in some decision situations but lead to absurd prescriptions in other cases. None of them can be accepted as the unique, completely convincing approach to be adopted. Hence, we have not succeeded in identifying a single substitute for the expected utility principle which cannot be applied under uncertainty and ignorance. Our collection of bridge-principles can at most count as a toolbox for decision making under uncertainty, with the flaw that we have hardly and only implicitly specified the conditions under which to apply this rather than another principle. It is as if we were plumbers, equipped with a rich tool-box but without having a concrete idea which of the tools to use when faced with a bursting of pipes. — Thus is the starting point of the next chapter. We have to understand, finally, what all this means for the way we take political decisions, for the involvement of scientific experts in democratic policy-making, for the role of science in a free society.

# Chapter 13

# Post-normal science


## Summary

The idea of post-normal science (PNS) as a way of managing uncertainties in science and politics was proposed by Funtowicz & Ravetz in the 1990s. Having defined PNS around the core-idea of lay-involvement, this chapter reconstructs two types of arguments in its favor: epistemic arguments on the one side and normative arguments on the other one. These arguments are primarily inspired by Beck, Feyerabend and Funtowicz & Ravetz themselves. Finally, this chapter suggests a slight modification of PNS in order to circumvent a potential problem.


## 13.1 Introducing post-normal science

Whereas the previous chapters stressed the challenge imperfect background knowledge represents for rational decision making, its implications are more ambivalent than it might have appeared so far. The following might be considered as a positive effect of uncertainty and ignorance: They are natural limitations to political power and control. Limits of science are at the same time limits to the "tyranny of science" (Feyerabend). Uncertainty and ignorance guarantee to some extent that one cannot control society the same way one controls a steam engine or a power plant. And they warrant that it is not only up to the cleverest scientists, the most ingenious technocrat to decide which policy to implement. Such were the predominant fears of Aldous Huxley, expressed for example by citing Tolstoi right at the beginning of *Science and Liberty* (Huxley, 1947):

> If the arrangement of society is bad (as ours is), and a small number of people have power over the majority and oppress it, every victory over Nature will inevitably serve only to increase that power and that oppression. This is what is actually happening.

Huxley, I suppose, thus would have welcomed our findings concerning the limits of science.

Yet, in the following we consider free societies where decision makers are not egoistic tyrants but elected and accountable representatives, and ask: How should policy-making be organized in fields characterized by uncertainty and ignorance? What role should scientific expertise play? Must science itself adapt to the new situation? — The paradigm of post-normal science (PNS) which will be discussed in this chapter addresses these questions.

What is PNS? PNS is a bunch of principles for policy-making under uncertainty and ignorance which was developed by Silvio Funtowicz and Jerome Ravetz in several articles in the 1990s. They have expressed the key ideas of PNS already in *Uncertainty and Quality in Science for Policy* (Funtowicz & Ravetz, 1990),

> We envisage a process of debate and dialogue operating continuously over all phases of a process, where uncertainty, quality and policy are inseparably connected. That the scientific inputs are contested along with the value considerations should be no occasion for alarm or dismay. This is what it is like, in the new world of policy-related research. Provided that debate is competent and disciplined, the only loss is of our rationalistic illusions. Beyond the inclusion of counter-expertise, we may imagine the diffusion of the requisite scientific skills among a broader population. (p. 67)

Accordingly, scientific statements have to be subjected to public critique and must not be taken for granted. This has to be done in an open debate "operating continuously over all phases" of the policy-making process. Whereas, under the traditional approach to decision making, the scientific *assessment* which yielded the required background knowledge (establishing the $\gamma$-tree) on the one hand and the *evaluation* of this knowledge including the decision-making based on some bridge-principle on the other hand could be and were separated, Funtowicz & Ravetz urge to merge assessment and evaluation. The critical discussion and rejection of this thesis will eventually, namely at the end of this chapter, lead to a modified proposal for policy-making under uncertainty and ignorance which is based on PNS.

The concept of PNS is explicitly introduced in *Science for the Post-Normal-Age* (Funtowicz & Ravetz, 1993),

> [Post-normal] science occurs when uncertainties are either of the epistemological or the ethical kind, or when decision stakes reflect conflicting purposes among stakeholders. [...] The problems are set and the solutions evaluated by the criteria of the broader communities. Thus

> post-normal science is indeed a type of science, and not merely politics or public participation. (pp. 749f.) ...
>
> The traditional fact/value distinction has not merely been inverted; in post-normal science the two categories cannot be realistically separated. The uncertainties go beyond those of the systems, to include ethics as well. (p. 751) ...
>
> The dynamic of resolution of policy issues in post-normal science involves the inclusion of an ever-growing set of legitimate participants in the process of quality assurance of the scientific inputs. (p. 752)

Thus, the principles expressed in Funtowicz & Ravetz (1990) are extended and complemented. Funtowicz & Ravetz propose that so-called extended peer communities which comprise all stakeholders set the standards and criteria for the relevant research activities. And an additional principle explicitly states that the traditional fact/value distinction has collapsed.

Yet, another year later, Funtowicz & Ravetz (1994) write,

> Here we will sketch the elements of a post-normal science [...]. These include the scientific management of uncertainty and of quality, the plurality of perspectives and commitments, and the intellectual and social structures that reflect the varied sorts of problem-solving activities. (p. 199)

The breakdown of the fact/value distinction has disappeared while the importance of management of uncertainty and quality is newly stressed. Extended peer communities can be considered as the appropriate "social structure" which reflects the "plurality of perspectives".

In a very recent article, Ravetz (2004) gives still another explication of PNS:

> [The] 'post-normal' approach embodies the precautionary principle. It depends on public debate, and involves an essential role for the 'extended peer community'. It is based on the recent recognition of the influence of values on all research [...]. It is the appropriate methodology when either systems uncertainties or decision stakes are high [...]. (p. XX)

He seizes most of the former principles but adds the precautionary principle as a defining characteristic of post-normal science.

PNS is, however, not only quite a loose concept but also defined by principles of a very different kind which actually depend on each other. Thus, the inseparability of facts and values seems to imply that stakeholders should be involved.

And extended peer communities are institutions that guarantee that the whole plurality of perspectives is taken into account. Given these diverse and interdependent aspects, I will narrow the definition to what I consider as the core of PNS. It will become clear during the discussion of the arguments for PNS how this core-definition is related to the remaining principles.

The core-definition reads: PNS refers to a policy-making procedure where decisions are not merely fueled by scientific expertise but evolve, right from the start of the research activities, out of an open debate in which scientists as well as laymen engage and where all statements are subjected to public critique. Extended peer communities are the institutional implementation of PNS.

The idea of PNS can be clarified further by contrasting it with its alternatives. Already the notion of post-normal science refers to what PNS is supposed not to be: normal science in the Kuhnian sense (Kuhn, 1962, pp. 23ff.). Hoyningen-Huene (1998) summarizes the characteristics of normal science as follows:

- the existence of regulations constraining acceptable approaches to and solutions of problems;
- the expectation of the solubility of appropriately chosen problems;
- no fundamental changes in the guiding regulations;
- absence of test or confirmation of the guiding regulations.

PNS as defined above obviously negates these points. Yet Funtowicz & Ravetz do not reject normal science in general but its application to specific policy problems, namely problems characterized by uncertainty and ignorance, as illustrated by figure 13.1: PNS is opposed to applied science and professional consultancy, both instantiating the idea of normal science. While under the paradigm of applied science decisions are based on scientifically established facts which are reviewed by scientific peer communities, under professional consultancy peer communities are typically composed of experts from different backgrounds. Each of these paradigms has its own scope of application but if uncertainty (represented by the x-axis) is high, policy-making should be organized in line with PNS.[1]

Although Funtowicz & Ravetz have introduced the notion of PNS, they were clearly not the first to urge a public control of and a stronger lay-involvement in science. Paul Feyerabend, for instance, vividly argued for the very same conclusion in *Science in A Free Society* (Feyerabend, 1978),

> [It] would not only be foolish *but downright irresponsible* to accept the judgement of scientists and physicians without further examination.

[1] The y-axis of figure 13.1 represents what Funtowicz & Ravetz consider as a further sufficient condition for applying PNS, i.e. the importance of the political decision to take.

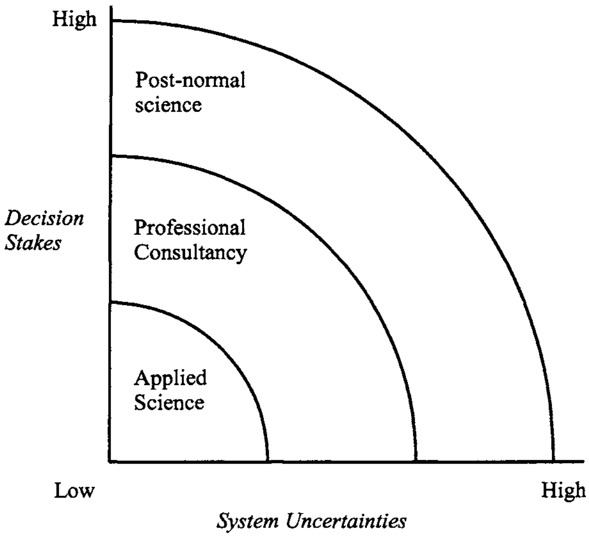



Figure 13.1: The scope of application of applied science, professional consultancy and post-normal science. (Adapted from: Ravetz, 2004)

> If the matter is important, either to a small group or to a society as a whole, *then this judgement must be subjected to the most painstaking scrutiny.* Duly elected committees of laymen must examine whether the theory of evolution is really as well established as biologists want us to believe, whether being established in their sense matters, and whether it should replace other views in school. They must examine the safety of nuclear reactors in each individual case and must be given access to all the relevant information. They must examine [...] (pp. 96f.)

Clarifying in a footnote that "the advice in all cases is to *use experts*, but never to *trust them* and certainly never to *rely on them* entirely", Feyerabend largely agrees with PNS.

Now that the paradigm of PNS is clarified, the question arises: Why should policy-making be based on it? Whereas there might be all kinds of arguments in favor of PNS, the following discussion is restricted to those which show that PNS should be applied under — and because of — uncertainty and ignorance. The arguments fall into two groups, namely epistemic and normative ones, and are mainly adapted from or inspired by three different parties: Paul Feyerabend, Ulrich Beck and Funtowicz & Ravetz & Co.[2]. Examples from Joseph Stiglitz's *Globalization and Its Discontents* (Stiglitz, 2002), a profound critique of the way

[2] "Co." like "commentators".

globalization is managed in general and of IMF policies in particular, will serve to illustrate the arguments and remind us what actually is at stake.

## 13.2 Epistemic arguments for PNS

The epistemic arguments for PNS warrant one and the same conclusion: (implementing) PNS tends to improve the quality of the background knowledge in general and to reduce the degree of ignorance in particular. Assuming that that is what we want, i.e. reducing ignorance, it follows that we should implement PNS.

*The argument from plurality of perspectives*

O'Connor (1999) expresses the general idea behind this argument,

> [...] many different points of view can be expressed, none of which is wholly convincing (to everybody, all of the time), none of which deals entirely adequately with all aspects of the situation, but none of which can be wholly rejected (by everybody) as having nothing at all relevant to say about the situation and about what should be done and why. (pp. 674f.)

So each person with its own individual perspective has something to add to the overall picture of a policy problem. The European Environmental Agency (2001) argues that that holds in particular for laymen,

> these may include industry workers, users of the technology and people who live in the locality or, because of their lifestyle or consumption habits, stand to be most strongly affected. [...]
>
> [The] benefit of attending to lay knowledge rests in its complementary character, its sometimes firmer grounding in real world operational conditions — as already discussed — and the associated independence from the narrow professional perspectives that can be a downside of specialist expertise. (pp. 177f.)

In terms of our general framework: The involvement of laymen — thanks to the complementary character of their knowledge — makes sure that further possible outcomes of a policy-measure which had not been taken into account before will be considered. The argument from plurality of perspectives can hence be reconstructed as follows,

(1) No single individual and no single epistemic community (defined by its shared background knowledge) is aware of all the aspects of a complex policy decision.

(2) The more different aspects are considered when exploring the $\gamma$-tree, the more possibilities will be taken into account and the more ignorance will be reduced.
(3) Extended peer communities represent the widest possible spectrum of different perspectives by recruiting scientists and (concerned and unconcerned) laymen.
(4) *Thus:* PNS reduces ignorance and improves the available background knowledge as far as possible.

While premiss (1) is an empirical statement, it holds for systematic reasons. Ongoing scientific progress (growth of knowledge-body) and limited human cognitive capacities require an ever narrowing specialization of scientists. This specialization inevitably leads to a very specific perspective which does not comprehend all the relevant aspects of a situation. An extension of this reasoning yields the argument from narrow-mindedness below.

The argument as reconstructed so far does not exclude other ways to enhance the background-knowledge. What about interdisciplinary expert committees? Maybe such very diverse groups including sociologists, anthropologists, biologists, economists, engineers, include as many different perspectives as a whole extended peer community. To strengthen the case for PNS, we thus have to point out that the perspectives of laymen are unique. Taking the above conclusion (4) as a starting point, the argument thence continues,

(4) PNS reduces ignorance and improves the available background knowledge as far as possible.
(5) Laymen, especially when directly concerned, have a unique perspective on a policy problem.
(6) Only PNS, relative to alternative ways of policy-making, guarantees the involvement of laymen in the assessment.
(7) *Thus:* None of the alternative ways of policy-making improves the available background knowledge as much as PNS does.

Laymen can improve our background knowledge by reducing uncertainty because of their unique perspectives. An outstanding example of a layman who has significantly reduced ignorance is Aldous Huxley. In his 1931 novel *Brave New World*, he depicted a future society where every aspect of the human condition is controlled and determined by means of applied science. Let us just consider one particular aspect of his scenario: the purposeful, industrial production of men as described in the first two chapters of the novel. In the Brave New World, men are not born but are the product of a sophisticated technological process. This process starts with the cloning ('bokanoskifying') of human eggs. The fertilized eggs are then filled into bottles where the growing embryo is submitted to dozens of different procedures aiming at conditioning it for its predestined place in society (for every individual's future purpose is known from the very beginning of the process as they are produced 'on demand' with a correspondingly large planning horizon). Yet, conditioning continues after the 'decantation' and young children are made "liking what they've got to do", as the director of the Central London Hatchery and Conditioning Centre puts it, by means of electro shocks and alarm bells.

It is rather difficult to grasp the novelty of Huxley's scenario today. Yet, the following note in a foreword to the novel written in 1946 might give us a rough idea about what creative and imaginative effort had been involved in depicting the Brave New World,

> In Brave New World this standardization of the human product has been pushed to fantastic, though not perhaps impossible, extremes. Technically and ideologically we are still a long way from bottled babies and Bokanovsky groups of semimorons. But by A.F. 600, who knows what may not be happening.

Thus even 15 years after having written the novel, Huxley considered the scenario still as being far from realizable. However, we did not even have to wait until A.F. 150, i.e. 150 years after the advent of Henry Ford, to see these possibilities becoming more and more feasible! Genetic engineering will possibly allow for much more precise, much more comprehensive and yet much more cost-efficient manipulation of men than Huxley originally imagined. Huxley's novel was not a prediction, it was the description of a possible scenario, a scenario nobody has seen as clearly as described by him before. This example underlines the importance to involve laymen, including artists, novelists, poets and other creative minds, in the scientific assessment of possible futures.

A less spectacular but no less important example are the IMF policies imposed on countries that suffer an economic crisis. Stiglitz (2002) notes that the IMF attitudes

> were antidemocratic. In our personal lives we would never follow ideas blindly without seeking alternative advice. Yet countries all over the world were instructed to do just that. (p. xiv)

Stiglitz, who was chairman of the Council of Economic Advisors under president Clinton, argues that the IMF should consult widely within a country because those "within the country are likely to know more about the economy than the IMF staffers—as I saw so clearly in the case of the United States" (p. 49). In fact, "if the IMF underestimated the risks to the poor of its development strategies, it also underestimated the long-term social and political costs of policies that devastated the middle-class, enriching a few at the top, and overestimated the benefits of its market fundamentalist policies" (p. 84). If laymen had been involved in assessing the possible outcomes of IMF policies, its possible implications for unemployment and poverty could not have been ignored and riots which frequently followed the implementation of IMF advice could have been prevented.

*The argument from narrow-mindedness*

This argument takes up a reasoning from the previous section. It argues that because scientists and experts are narrow-minded, they tend to make mistakes by omitting possibly important aspects. Since such mistakes can be corrected by laymen, their involvement improves our background knowledge.

One reason for the scientists' narrow-mindedness is given by Beck (1986),

> It is not the failure of individual scientists or disciplines, but it is systematically rooted in the institutional-methodological approach of the sciences to risks. The sciences as they are organized today — with their highly specialized division of labor, with their understanding of methods and theories, with their externally dictated abstinence from practical problems — are simply not in a position to adequately respond to modern risks ["Zivilisationsrisiken"]. (p. 78)

According to Beck, it is the highly specialized division of labor which necessarily leads to narrowed perspectives. Feyerabend (1978) underpins this sociological observation with a general methodological argument,

> Every piece of knowledge contains valuable ingredients side by side with ideas that prevent the discovery of new things. Such ideas are not simply errors. They are necessary for research: progress in one direction cannot be achieved without blocking progress in another. But research in that 'other' direction might reveal that 'progress' achieved so far is but a chimera. (p. 89)

But can the 'mistakes' of highly specialized scientists which consist in omitting possibly important side-effects, or in not pursuing worthwhile alternative directions of research be corrected by laymen? Feyerabend strongly believes so,

> That the errors of specialists can be discovered by ordinary people provided they are prepared to 'do some hard work' is the basic assumption of any trial by jury. [...] This assumption is confirmed in trial after trial. [... Science] *is not beyond the reach of the natural shrewdness of*

*the human race*. I suggest that this shrewdness be applied to all important social matters which are now in the hands of experts. (p. 97)

and he gives numerous examples of 'outsiders' and self-proclaimed 'dilettantes' who made valuable contributions to science such as Einstein, Bohr or Born to name just three prominent ones (see Feyerabend, 1978, p. 88). This said, the argument of narrow-mindedness can concisely be written as,

(1) Scientists are (for systematic, methodological reasons) highly specialized and thus narrow-minded.
(2) Narrow-mindedness might lead to mistakes due to omission.
(3) Laymen can correct mistakes which result from narrow-mindedness.
(4) In extended peer communities, layman supervise and correct scientists.
(5) *Thus:* PNS improves the available background knowledge.

According to Stiglitz, narrow-mindedness is one of the main problems of the IMF: Its staff — in particular its executive managers — consist of specialized economists who believe that economic problems can be solved by considering economic variables and applying economic policies only. But Stiglitz (2002) stresses,

> The very notion that one could separate economics from politics, or a broader understanding of society, illustrated a narrowness of perspective. (p. 47)

Had the IMF research activities and strategy formulation been reviewed by an extended peer community, such narrow-mindedness could have been corrected.

*The argument from unanimity*

We reconstruct the argument from unanimity by Feyerabend (1978, pp. 88f.). Its first part is a classical dilemma and warrants the conclusion that scientists tend to make mistakes: Scientists either disagree on a policy recommendation or propose it unanimously. If they disagree, at least one of them makes a mistake. Unanimity, on the other hand,

> is often the result of a *political* decision: dissenters are suppressed, or remain silent to preserve the reputation of science as a source of trustworthy and almost infallible knowledge. On other occasions unanimity is the result of shared prejudices [...]. Then again unanimity may indicate a decrease of critical consciousness [...]. (Feyerabend, 1978, p. 88)

And all these attitudes are potential sources of mistakes, too. Assuming, similar to the argument of narrow-mindedness, that such mistakes can be corrected by laymen, we can conclude that PNS improves the available background knowledge. In brief,

(1) If scientists disagree on an issue, at least one of them makes a mistake.
(2) Unanimity of scientists, on the other hand, is sometimes the result of suppressing dissenters or shared prejudices.
(3) A recommendation which was reached by suppressing dissenters or which is based on shared prejudices might be mistaken.
(4) Laymen can correct mistakes, in particular those which arise out of prejudices or because dissenters are suppressed.
(5) In extended peer communities, laymen supervise and correct scientists.
(6) *Thus:* PNS improves the available background knowledge.

Again, Stiglitz (2002, p. 220) provides a nice illustration. He notes that there is some general disagreement among economists as to which strategies are best suited to combat (financial) economic crises (first horn of the dilemma). However, according to Stiglitz's account, the IMF presents its particular advice as the one and only truth and as if there were no scientific controversy on these issues at all, ignoring its discontents (second horn of the dilemma). Thus, global economic policy-making will improve from lay-involvement.

The argument from unanimity, however, faces the following critique. Statements and recommendations which are based on prejudices, which arose out of a narrowed point of view or which were not submitted to open critique certainly lack credibility and policies should not be based on them. Still, these are clearly not the virtues we traditionally attribute to science — but their reverse. The ideal of the scientific method includes open, unprejudiced debate and critique. In this light, Feyerabend's arguments may be read as a plea for more good science rather than for control of science. Instead of introducing external supervision of scientific activities, we should foster the traditional virtues of science. But these are no exclusive alternatives, as the next argument will show. In fact, PNS is an appropriate way to compensate the systematic deviation of scientific practice from its ideal.

*The argument from self-criticism*

The main idea of this argument is that PNS enables scientists to exercise effective self-criticism — a capacity they have lost at least to some extent because of conflicts of interests. Beck (1986) stresses the need to ensure, by the purposive design of institutions, that science not only controls but also criticizes itself,

> The possibility of self-control, praised by all monopolists, must be complemented by the possibility of *self-critique*. That is: What had to fight its way against the dominance of professions or operational management so far must be secured institutionally: counter-expertise, alternative professional practice, inner-professional and -organizational controversies about the risks of developments, suppressed skepticism. In this case Popper is right indeed: Critique means progress. Only where medicine stands against medicine, nuclear physics against nuclear physics, human genetics against human genetics one can see and judge from outside the future that is in the making. Enabling self-critique of all kinds is by no means a danger but the truly only way how the error which will otherwise blow the world to pieces sooner or even more sooner could be detected in advance. (p. 372)

Contrary to Healy (1999) who considers Beck's proposal to strengthen the scientists' capacity of self-criticism as an alternative to PNS, I believe that the two can be reconciled. All the more as Beck's ideas even give rise to an argument in favor of PNS. For indeed, extended peer communities are an appropriate institutional set-up which reintroduces and guarantees effective self-criticism of science: Dissenters who might have been ignored inside their scientific community will be heard in extended peer communities, mainstream scientists will be forced to deal with non-orthodox critiques. Because PNS recognizes the plurality of perspectives, medicine will stand against medicine and nuclear physics will stand against nuclear physics in extended peer communities. These communities can therefore be regarded as blend-institutions which guarantee that the scientific mainstream doesn't directly, i.e. without being submitted to critique, without being blended with other points of view, fuel the policy making process. The argument from self-criticism hence becomes,

(1) Scientific research and progress is frequently the initial cause of new dangers and risks and therefore needs to be regulated.
(2) *Thus:* Scientists, when advising policy-makers, sometimes have to assess their own profession.
(3) The institutional organization of western science tends to punish dissenters and self-critique and to promote conformity.
(4) *Thus:* Scientific advisors face a conflict of interest and their assessments are sometimes likely to be fudged.

(5) Extended peer communities correct such assessments, decrease the pressure for conformity and enable scientists to exercise self-critique.
(6) *Thus:* PNS improves the available background knowledge.

As might have been expected from the previous examples, the IMF seems to be missing the capacity to criticize itself, too. Stiglitz (2002) says that "the last thing [the World Bank and the IMF] wanted was a lively democratic debate about alternative strategies" (p. 15). And the IMF "believed it [its market doctrine] so strongly that it did not need to look at any evidence and gave little credence to any evidence that suggested otherwise" (p. 208). The IMF's unwillingness to effectively criticize itself also became apparent in the reaction to Stiglitz's book (see Stiglitz, 2002, pp. 272f.).

*The argument from tacit knowledge*

When reconstructing the argument from plurality of perspectives, we have assumed that the specific perspective of laymen is unique. That has been the reason why they should be involved in the scientific assessment of a policy situation. Now why can't scientists acquire this unique knowledge by interviewing experienced laymen? This would make lay-involvement — at least insofar as justified by the plurality argument — obsolete.

However, knowledge that has been acquired by long lasting experience is frequently tacit. And as such, it cannot be 'extracted' out of the lay person. In addition, it is this tacit knowledge which matters during the scientific assessment because informed judgment is required when deciding which possibilities to include in the $\gamma$-tree and which not. Such decisions cannot rely solely on rules or fixed standards as Funtowicz & Ravetz (1990) remarked generally,

> The set of explicit rules can be insufficient for the proper operation of the system. Judgements must be made for interpreting or occasionally even over-riding the rules. These depend on the skills of recognition of the degree to which a given situation matches that in a rule-book, and also the skills of operating in circumstances not envisaged in the rule-book. In the absence of such skills, either blind obedience or random panic reactions are the only alternatives. (pp. 40f.)

The following argument summarizes this reasoning by restating the argument from plurality,

(1) Tacit knowledge which has been acquired through long lasting experience and acquaintance with the corresponding system might be

crucial when deciding which possibilities to include or to exclude in the $\gamma$-tree, i.e. for reducing uncertainty and ignorance.

(2) Scientists and experts have no more, and often less, of such tacit knowledge than practitioners, experienced laymen and decision-makers.

(3) Extended peer communities and the open dialogue make sure that the implicit knowledge of such persons will be considered, too.

(4) *Thus:* PNS improves the available background knowledge.

*The argument from plurality of scientific traditions*

The argument from narrow-mindedness stressed that scientists are highly specialized and that the limits of their point of view incline to coincide with the limits of their sub-discipline. This idea can be generalized as western scientists tend to be narrow-minded not only with respect to their sub-discipline, but also with respect to the scientific tradition they belong to.

Feyerabend (1978, pp. 100ff.) notes that there is no competition between different scientific traditions today and that western science is the only universally accepted one. Yet, this does not simply stem from its merits.

> It reigns supreme because some past successes have led to institutional measures (education; role of experts; role of power groups such as the AMA) that prevent a comeback of the rivals. Briefly, but not incorrectly: today science prevails not because of its comparative merits, but because the show has been rigged in its favor. (p. 102)

For other scientific traditions might very well have something to add to our knowledge, it is important to reinstall the competition between different scientific traditions.

This reasoning is strengthened by Beck (1986) who observes,

> The currently accepted knowledge of the risks and dangers of the scientific-technological civilization had actually to be established against the massive denials, against the frequently fierce resistance of a 'scientific-technical rationality' that has been self-sufficient and stubbornly prejudiced due to its faith in progress. The scientific risk-research limps behind the socio-environmental, progress related and cultural critique of the industrial system in every respect. (p. 77)

Hence, there are not only alternatives to western scientific tradition, but this very tradition has so far and in some cases prevented us from acquiring crucial knowledge of future dangers and risks. Thus a further reason to complement it by alternative approaches.

The fact that extended peer communities include representatives of alternative scientific traditions and therefore foster the co-existence and the healthy competition of scientific cultures completes the argument.

(1) Western science is by no means the only scientific tradition which might give rise to valuable knowledge.
(2) Western science has been denying and playing down much of the dangers of modern technology we are nowadays aware off.
(3) *Thus:* The competition and mutual supplementation of scientific traditions will enhance our background knowledge.
(4) Extended peer communities consider different scientific traditions and foster their co-existence.
(5) *Thus:* PNS improves the available background knowledge.

As an example, consider the IMF again. Stiglitz (2002) criticizes the IMF policies for being based on a very specific tradition of economic science,

> Decisions were made on the basis of what seemed a curious blend of ideology and bad economics, dogma that sometimes seemed to be thinly veiling special interests. (p. xiii)

This would be impossible if extended peer communities assessed and discussed the economic crises the IMF is supposed to deal with.

## 13.3 Normative arguments for PNS

In contrast to the epistemic arguments, the normative arguments for PNS do not only warrant the same conclusion but also share a common assumption. This is the assumption that what is normatively right or wrong is not objectively given, i.e. moral realism is rejected. Accordingly, there is no expert knowledge in the domain of ethics, and the right values are those which are agreed upon intersubjectively. Generally, this assumption endorses some kind of participatory conception of ethics as discourse ethics. In terms of political philosophy, it urges that decisions must be taken democratically (for normative reasons).

Sharing this assumption, the normative arguments try to show that PNS is the only way to guarantee that all those who are concerned participate in every decision involving a value choice.

*The argument from limits of individual rationality*

We have found in the previous chapter that there is no unique bridge-principle for decision making under uncertainty and ignorance. In contrast, several alternative bridge-principles might be equally appropriate for one and the same decision problem. Thus the choice of the bridge-principle to adopt is to some extent arbitrary and reflects some underlying values. Furthermore, once a bridge-principle is agreed upon, there might still be several different strategies which are compatible with the bridge-principle and the background knowledge so that yet another arbitrary choice has to be made. It is this underdetermination Jaeger et al. (1998) express in different terms,

> The problem with the attempt to approach risk-studies through the establishment of one privileged expert consensus lies in the possibility that more than one view of a specific risk may be perfectly defensible. Although this possibility runs against the thrust of traditional philosophy of science, it is easy to understand even from the perspective of the rational actor paradigm. If the definition of risk is a matter of utility functions and subjective probabilities, there is no reason to expect a single definition to be superior to all other ones. Perfectly rational people may hold different utility functions and subjective probabilities [...]. (p. 185)

Now, given (i) the normative premiss that those who are concerned should participate or at least be represented in a value choice as well as (ii) the fact that any such participation is *per definitionem* an extended peer community, we may infer that policy making under uncertainty and ignorance should correspond to the principles of PNS.

(1) For any decision-problem under uncertainty and ignorance, there is more than one appropriate bridge-principle.
(2) A choice between equally justified alternatives expresses some values and preferences.
(3) *Thus:* Adopting a specific bridge-principle is a choice which expresses some underlying values.
(4) Value choices should be made by those who are concerned.
(5) Extended peer communities and extended peer communities only guarantee that all those who are concerned participate or are represented in the choice for a certain bridge-principle.
(6) *Thus:* Under uncertainty and ignorance, decisions should be taken in line with the principles of PNS.

This argument urges that extended peer communities should carry out the *evaluation* — in addition to the scientific assessment — of the $\gamma$-tree, and take the final decision. This, however does not distinguish between decision making under uncertainty on the one hand and under certainty on the other. For even under certainty, the evaluation of the $\alpha$-tree requires a preference-relation among the possible outcomes which is nothing but a normative judgment. The evaluation of a policy problem thus should always, and not only under uncertainty and ignorance, be carried out by an extended peer community or an other democratic institution which ensures the involvement of all who are concerned.

*The argument from value-ladenness*

As we have already seen when introducing PNS, the idea of value-ladenness is at its heart. Consider again the quote from Funtowicz & Ravetz (1993),

> The traditional fact/value distinction has not merely been inverted; in post-normal science the two categories cannot be realistically separated. (p. 751)

According to my interpretation of PNS, this statement does, however, not define PNS but maintains that facts are inevitably value-laden under sufficiently large uncertainty and ignorance. A similar observation is made by Beck (1986),

> The claim to assess the extent of a certain risk in a rational and objective way permanently disproves itself: [...] In addition, one must have adopted a certain normative stance in order to be able to talk meaningfully about risk at all. (pp. 38f.)

Therefrom Beck may infer that,

> By dealing with risks the natural sciences have unintentionally and to some extent *deprived themselves of their power*, have compelled themselves to become more *democratic*. (p. 77)

So far, the inevitable value-ladenness of scientific facts under uncertainty is merely a claim and not warranted yet. Do Funtowicz & Ravetz argue for this very thesis? Well, there is one line of argument which says that in addition to factual uncertainties PNS has to deal with real ethical uncertainties, too, as "it is impossible to produce a simple rationale for adjudicating between the rights of people who would benefit from some development, and those of a species of animal or plant which would be harmed" (Funtowicz & Ravetz, 1993, p. 751). But even if Funtowicz & Ravetz were able to show that the emergence of factual and ethical uncertainties coincides, this would not at all yield the desired conclusion of value-ladenness.

Yet, Ravetz (2004) seems to provide a compelling example for the inevitable value-ladenness of facts,

> After many generations of propaganda for science proclaiming its freedom from values, the secret has leaked out that all statistical tests are value-loaded, necessarily designed to avoid one or another sort of error [...]. Without going into technicalities, we may say that any test might be overly selective, rejecting correlations that are probably real; or it might be overly sensitive, accepting correlations that are probably accidental. This distinction is most frequently expressed through a 'confidence limit', where a high confidence limit protects against oversensitivity but makes the test vulnerable to over-selectivity. What is appropriate for a laboratory experiment, where the main concern is protecting the research literature from spurious results, may be quite inappropriate for exploratory or monitoring research, where weak signals of harm may be all that we have. It is impossible to design a statistical test which avoids both types of error; there must be a choice, made by someone, somewhere. (p. 5)

Before discussing this issue in further detail, let us have a look at the whole argument in its concise version and consider an example.

(1) There is no value-free risk assessment. In complex decision-situations, almost any scientific fact holds conditional to some implicit normative assumptions.
(2) *Thus:* The acceptance of a fact requires the acceptance of some specific values.
(3) Value choices should be made by those who are concerned.
(4) *Thus:* Which facts are to be accepted may not be decided by scientific experts alone.
(5) Extended peer communities and extended peer communities only guarantee that all those who are concerned participate or are represented in which facts to accept.
(6) *Thus:* Under uncertainty and ignorance, decisions should be taken in line with the principles of PNS.

Value-ladenness is not merely an abstract problem as the IMF case illustrates. Stiglitz (2002) points to the value-ladenness of IMF policies by noting that the "lending decisions were political—and political judgment often entered into IMF

UNCERTAINTY-ASSESSMENT REPORT NO. 1
prepared for the Committee ...
by the Task Force for the Assessment of Policy-Options.
*Blueprint*

*Task:* We have been charged by the committee to set up alternative policy strategies for the handling of ... and assess their possible outcomes.

*Strategies:* We have examined the following range of policies: $s_1, \ldots, s_n$.

*Possible outcomes:* The assessment involved several arbitrary choices so that the results depend on the underlying assumption-cluster.

- Assumption cluster A gives rise to the following $\gamma$-tree: ...
- Assumption cluster B gives rise to the following $\gamma$-tree: ...
- Assumption cluster C gives rise to the following $\gamma$-tree: ...
- ...

*Caveat:* We would like to remind the committee that the above $\gamma$-trees might be incomplete for there might be possible outcomes we totally ignore and which are not included in this report.

Figure 13.2: Blueprint for an uncertainty-assessment report.

advice" (p. 47). He adds that "the lack of concern about the poor was not just a matter of views of markets and government, views that said that market would take care of everything and government would only make matters worse; it was also a matter of values—how concerned we should be about the poor and who should bear what risks" (pp. 85f.). But instead of giving a priority to poverty reduction, "U.S. economic interests—or more accurately, U.S. financial and commercial market interests were reflected in the policies" (pp. 171f.).

However, to return to the abstract argument, I maintain that premiss (1) is not generally true and although 'scientific' recommendations are sometimes value-laden, this does not *have* to be the case. I will argue that the separation of facts and values can principally be achieved in every situation. Thus, recall first that the $\gamma$-tree displays modal, non-normative information only. If a scientific assessment which gives rise to a $\gamma$-tree is therefore value-laden, some of the statements represented by the $\gamma$-tree have to hold conditional to some normative assumptions, e.g. they are based on some arbitrary choice. Now if such is the case, if the $\gamma$-tree inevitably rests on some arbitrary choices, then the assessment should consists of several different $\gamma$-trees corresponding to the alternative choices. Assume for example that some information incorporated in a $\gamma$-tree is based on the statistical analysis of some data. This analysis should then be accomplished for different confidence intervals and if these yield different results, the scientific assessment should include several $\gamma$-trees, namely one for each confidence interval. The blueprint for an uncertainty-assessment report shown in figure 13.2 summarizes the ideas outlined in this paragraph.

If the above guidelines are followed, value-ladenness can be avoided during the uncertainty assessment. The argument from value-ladenness does consequently not provide a reason to set up extended peer communities.

*A pessimistic induction and the argument from trust*

We have just seen that the separation of value-free assessment and normative evaluation is possible at least in principle. However, Jaeger et al. (1998) note that this has rarely been achieved,

> In many cases, especially in the areas where risk management based on the rational actor paradigm has been most conspicuously applied, the separation of assessment and evaluation seems to have been ineffective and, at times, counterproductive, leading to more opposition and mistrust over time, instead of less. (p. 174)

These observations might give rise to an inductive, empirical and rather pragmatic argument against the separation of assessment and evaluation and for PNS.

(1) As a matter of fact, the separation of risk assessment and evaluation has rarely been effective.
(2) *Thus:* The separation of risk assessment and evaluation is not effective.
(3) If assessment and evaluation cannot be effectively separated, one should not attempt to do so.
(4) Only PNS appropriately merges risk assessment and evaluation.
(5) *Thus:* Under uncertainty and ignorance, decisions should be taken in line with the principles of PNS.

The fact that attempts to separate assessment and evaluation have not been effective yet, poses also a more general problem for the legitimation of government, as Marchi & Ravetz (1999) see,

> These new hazards combine extreme uncertainty with the possibility of extended and irreversible harm. For them new forms of governance will be required. In the absence of reliable scientific evidence, governance in such difficult cases becomes more dependent on the trustworthiness

> of the authorities, achieved partly by an extension of public participation. (p. 755)

The argument from trust hence becomes,

(1) The legitimation (as opposed to legitimacy) of government action depends crucially on its trustworthiness.
(2) Under uncertainty and ignorance, i.e. under the absence of clear solutions to policy problems, trustworthiness cannot be achieved by reference to scientific expertise. (As attempts to separate risk assessment and evaluation typically result in mistrust.)
(3) Yet, participation is an effective means to foster the citizens' trust in their government.
(4) *Thus:* Participation becomes an important condition for legitimate government.
(5) PNS is an appropriate framework for largest possible participation.
(6) *Thus:* Under uncertainty and ignorance, decisions should be taken in line with the principles of PNS in order to enhance government legitimation.

Both arguments are based on the inductive inference which concludes that assessment and evaluation cannot be effectively separated. Still, this inference is fallacious. The failure of attempts to separate assessment and evaluation only shows that the above guidelines have not been applied strictly. Doing so will help to restore trustworthiness not only of government action but also of scientific expertise.

*The argument from entanglement*

So far, we have rejected one major argument which proclaimed the breakdown of the traditional fact-value separation. Here is yet a much more cautious critique of this distinction, put forward by Hilary Putnam who agrees that facts can be distinguished from value-statements. He simply stresses that there is no sharp borderline between the two and that

> conceding that there is a class of (paradigmatically ethical) judgements that contain perhaps nine or ten or a dozen familiar ethical words neither solves any philosophical problems, nor tells us what exactly makes a word an ethical word, nor requires us to concede that all the non-ethical judgements fall into one or even two or three natural kinds. (Putnam, 2002, p. 16)

According to Putnam and several other philosophers, there are many so-called thick concepts which are neither purely normative nor exclusively descriptive but blended. The inseparability of facts and values accordingly becomes a linguistic thesis: Values and facts are entangled because the meaning of many words comprehends a normative as well as a descriptive component.

This kind of entanglement could pose a problem for value-free uncertainty-assessment if the description of the $\gamma$-tree, i.e. of the alternative strategies or the possible outcomes, necessarily involved thick concepts. For even describing a scenario in these rather than in other words would then be an arbitrary choice reflecting some underlying value-statements. This reasoning results in the following argument for PNS,

(1) The description of possible outcomes will necessarily involve thick concepts.
(2) The meaning of thick concepts includes a normative component.
(3) *Thus:* Adopting a specific description implies to accept certain value statements.
(4) Value choices should be made by those who are concerned.
(5) *Thus:* Which descriptions are to be adopted may not be decided by scientific experts alone.
(6) Extended peer communities and extended peer communities only guarantee that all those who are concerned participate or are represented in which descriptions to adopt.
(7) *Thus:* Under uncertainty and ignorance, decisions should be taken in line with the principles of PNS.

Now, as a matter of fact, thick terms like 'acceptable risk', 'severe risk', 'catastrophic damage', ... seem to be used in scenario descriptions. But does this have to be the case? Are thick concepts indispensable for depicting possible outcomes? Are they irreducible? Or can we always give a neutral, non-normative account of a possible outcome? Answering these questions is beyond the scope of our enquiry as it would require a couple of detailed case studies to get an idea about which languages are appropriate for scenario based planning and which are not.[3]

Insofar as possible outcomes can really not be described in a value-neutral language, we have here a normative argument for lay-involvement in the uncertainty-assessment. Even then, the separation of facts and values, although impossible to realize fully, should remain a regulative ideal of the policy-process.

[3] Thereby, the family of IPCC scenarios would definitively be an interesting case to start with.

## 13.4 Dangers and a modification of PNS

Though there are several reasons to organize policy making according to the principles of PNS, we may not omit the risks of doing so. Just imagine a debate where particle physicists as well as plumbers, post-modern sociologists as well as train conductors, lawyers and homeopaths, politicians and astrologers participate all alike. Will this debate ever lead to some result? If every (scientific) standard is contested in PNS, by which criteria is quality defined in PNS? Based on which grounds will the extended peer community reach some agreement if there is a plurality of scientific cultures and epistemic traditions?

However, the risk is much larger than a simple deadlock of policy-debates for PNS is supposed to be not only a procedure for policy-making but a way to reorganize science in policy-relevant fields. Yet, will science be able to fulfill one of its important social functions, namely to provide means to question and criticize social and political practices, if standards of quality are qualified and science is at least a priori not given more credit than any other cultural enterprise? Should we brush off the advice that a member of the Constituent Assembly of the French Revolution gave his co-revolutionaries?

> Let us not forget that long before we did, the sciences and philosophy fought against the tyrants. Their constant efforts have made the revolution. As free and grateful men, we ought to establish them among us and cherish them for ever. For sciences and philosophy will maintain the liberty which we have conquered. (quoted from Hobsbawm, 1977, p. 336)

Beck (1986) pinpoints the current risk of immunizing social practices against any critique by relativizing science,

> [This development] might give rise to an emancipation of social practices *from* science *through* science; on the other hand, it *immunizes* accepted ideologies and interests against scientific critique and opens the floodgates to a feudalization of scientific authority by economic-political interests and 'new ideological powers'. (p. 257)

Accordingly, not only will it become more and more difficult to criticize some social practice if the plurality of scientific traditions is fully endorsed but also it will become easier to 'scientifically' legitimate any practice.

These risks are to some extent addressed by Ravetz (2004),

> There is a systematic problem here: if traditional research science required collegiality and a shared idealism for its maintenance, how can

> quality survive in the adversarial atmosphere of policy-relevant science. [...] To be sure, when violent debate is dominant, then accusations of 'junk science' will fly in all directions. But the process of post-normal science is one of mutual learning. It first requires respect and then appreciation, of the perspectives and commitments of other parties in the extended peer community. Trust can be built, and then extended peer review can be done successfully. (p. 10)

More precisely, what guarantees the quality of information and background knowledge in PNS is *the quality of the debate* in the extended peer communities. It should conform as much as possible to a non-authoritarian, argumentative discourse between equal participants who thereby engage in "communicative action" (Habermas). If debate in the extended peer communities conforms to this ideal at least partly, then a society will improve and not degrade its overall capacity to deliberate and criticize.

As a result of the discussion of PNS in the course of this chapter I propose to modify the idea of PNS slightly. Under uncertainty and ignorance, the whole political process should be divided into two steps. The first step consists in the assessment of the options and their possible outcomes which eventually gives rise to a (or several) $\gamma$-tree(s). For *epistemic* reasons, this assessment should be accomplished by extended peer communities. By trying to avoid thick concepts, they can ensure that the $\gamma$-tree is nearly value-free. The second step comprises the evaluation of the $\gamma$-tree, including the choice of a bridge-principle and the final decision. This evaluation must be accomplished by extended peer communities or other institutions which guarantee the involvement of all who are concerned for *normative* reasons. In sum, this modification structures the political process and separates different debates so as to increase transparency, minimize complexity and avoid possible confusions.

The examples taken from Stiglitz's critique of global politics should have made clear: The ideas discussed in this chapter and which conclude our investigation are not merely of academic interest but have a huge potential impact on how we shape our common future. If we stop pretending to have valuable foreknowledge and if we fully acknowledge our ignorance, we will be able to take more reasonable, more honest, and more democratic decisions. Yet, one cannot underestimate the challenge this represents: Switching the mode of our debates from perfect foresight and risk to uncertainty and ignorance, and from scientific authority to pluralism will multiply the complexity of argumentation. It is an open question whether this will exceed our collective (institutional) and individual (cognitive) capacities.

# Appendix

## A: Taylor diagram

The following reasoning shows that the linear distance between a point representing a model and the point of reference in a Taylor diagram (figure 4.10) represents the part of the RMSE that is not due to bias, divided by the observed variable's standard deviation. Let us consider the forecast ${}_{t-1}X_t^F$ as random variable, then the forecast error $e_t = {}_{t-1}X_t^F - X_t$ is a random variable, too, and as we have seen in Chapter 2, the MSE can be written as

$$\mathrm{MSE} = \mathrm{E}((e_t)^2)$$

Substitution that makes use of the definition of variance and covariance yields

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}(({}_{t-1}X_t^F - X_t)^2) = \mathrm{E}(({}_{t-1}X_t^F)^2 - 2{}_{t-1}X_t^F X_t + (X_t)^2) \\
&= \mathrm{E}(({}_{t-1}X_t^F)^2) - 2\mathrm{E}(({}_{t-1}X_t^F)X_t) + \mathrm{E}((X_t)^2) \\
&= \mathrm{V}({}_{t-1}X_t^F) + \mathrm{E}({}_{t-1}X_t^F)^2 + \mathrm{V}(X_t) + \mathrm{E}(X_t)^2 \\
&\quad -2COV({}_{t-1}X_t^F, X_t) - 2\mathrm{E}({}_{t-1}X_t^F)\mathrm{E}(X_t) \\
&= \underbrace{(\mathrm{V}({}_{t-1}X_t^F) + \mathrm{V}(X_t) - 2\mathrm{COV}({}_{t-1}X_t^F, X_t))}_{\mathrm{MSE_{var}}} \\
&\quad +(\mathrm{E}({}_{t-1}X_t^F)^2 + \mathrm{E}(X_t)^2 - 2\mathrm{E}({}_{t-1}X_t^F)\mathrm{E}(X_t))
\end{aligned}$$

This expression nicely pinpoints the two components of the MSE: The first bracket represents the part that stems from the variables' volatility, and the second bracket stands for the error solely due to bias. It follows for $\mathrm{MSE_{var}}$

$$\begin{aligned}
\frac{\mathrm{MSE_{var}}}{\mathrm{V}(X_t)} &= \frac{1}{\mathrm{V}(X_t)}(\mathrm{V}({}_{t-1}X_t^F) + \mathrm{V}(X_t) - 2\mathrm{COV}({}_{t-1}X_t^F, X_t)) \\
&= 1 + \frac{\mathrm{V}({}_{t-1}X_t^F)}{\mathrm{V}(X_t)} + \frac{2\mathrm{COV}({}_{t-1}X_t^F, X_t)}{\mathrm{V}(X_t)} \\
&= 1 + \frac{\mathrm{V}({}_{t-1}X_t^F)}{\mathrm{V}(X_t)} + 2\frac{\sqrt{\mathrm{V}({}_{t-1}X_t^F)}}{\sqrt{\mathrm{V}(X_t)}}\frac{\mathrm{COV}({}_{t-1}X_t^F, X_t)}{\sqrt{\mathrm{V}({}_{t-1}X_t^F)\mathrm{V}(X_t)}}
\end{aligned}$$

$$= 1 + \frac{\mathrm{V}({}_{t-1}X_t^F)}{\mathrm{V}(X_t)} + 2\frac{\sqrt{\mathrm{V}({}_{t-1}X_t^F)}}{\sqrt{\mathrm{V}(X_t)}} R_{{}_{t-1}X_t^F X_t}$$

where $R_{{}_{t-1}X_t^F X_t}$ is the correlation coefficient between the predicted and the observed values. Now, the special scale of the half circle in the Taylor diagram ensures that $R_{{}_{t-1}X_t^F X_t} = \sin(\alpha)$ for a given forecast represented in the diagram with $\alpha$ being the angle $\langle$forecast-point–(0,0)–point of reference$\rangle$. According to pythagoras and basic trigonometry it follows that the linear distance between the forecast-point and the point of reference is equal to the RMSE not due to bias normalized by the standard deviation $\sigma$ of the observations, namely $\frac{\mathrm{RMSE}_{var}}{\sigma} := \sqrt{\frac{\mathrm{MSE}_{var}}{V(X_t)}}$. Thus, if the forecast is not biased, the linear distance represents the relative RMSE adjusted for the predicted variable's volatility.

## B: Original quotes

Page 83:

> Der Satz: Wirtschaftsprognose ist mit den Hilfen, die ökonomische Theorie und Wirtschaftsstatistik bieten, grundsätzlich unmöglich, [...] muß [...] dahingehend interpretiert werden: Auch wenn es eine positive Theorie der Wirtschaftsprognose gäbe, was wir leugnen, so wäre Wirtschaftsprognose als solche noch immer ungenügend, weil die Wirtschaftsabläufe durch andere Umstände, andere Verhaltensarten mitbedingt sind, die gleichfalls mitvorausgesagt werden müßten, aber nicht vorausgesagt werden können. Oder anders: Eine Prognose von wirtschaftlichen Vorgängen ist *eo ipso* eine Mitprognose von anderen Verhaltensarten, die sich im wirtschaftlichen Verhalten irgendwie quantitativ niederschlagen. Das geht aber unter allen Umständen über den Möglichkeitsbereich der ökonomischen Theorie hinaus, ganz gleichgültig, ob die Theorie der Prognose noch so brilliant sein mag oder nicht.

Page 104:

> Keineswegs alle Datensätze gehen aus direkten und auf diesen Erkenntniszweck bezogenen Erhebungen hervor. Viele Informationen müssen durch Kombination von Daten gewonnen werden, die für diese Kombination ursprünglich nicht gemacht sind. Viele Daten müssen auch durch äußerst kreative Schätzungen produziert werden. Man wird wohl auch heute mit einer Mischung aus Respekt und leiser Bangigkeit sagen dürfen: In den VGR [Volkswirtschaftlichen Gesamtrechnungen] muss auch ein wenig gezaubert werden.

Page 105:

Das Aufstellen von Volkswirtschaftlichen Gesamtrechnungen ist angesichts der Unvollkommenheit der statistischen Fundierung eher als kreativer Vorgang denn als Rechenprozess zu verstehen.

Page 105:

Welcher Wert innerhalb der Bandbreite der vom Schätzenden nach Ausschöpfung aller Informationen als gleich glaubwürdig beurteilten Ergebnisse für den einzelnen Tatbestand letztendlich veröffentlicht wird, ist grundsätzlich ohne Bedeutung und in der Tat mehr oder weniger willkürlich.

Page 105:

Den wahren Wert des Bruttoinlandsprodukts kennt niemand. Die vom Statistischen Bundesamt für Deutschland veröffentlichten Zahlen stellen jeweils die aus Sicht der für die Berechnung und Veröffentlichung Zuständigen 'beste' Zahl dar. Sie ist im Prinzip immer 'vorläufig'.

Page 105:

Der Datennutzer ist stets in Gefahr, Regelmäßigkeiten, Beziehungen, strukturelle Stabilitäten *herauszurechnen*, die das statistische Amt bei der Schätzung der Daten *hineingerechnet* hat, weil nämlich nur mittels Verhältnisschätzungen oder Regressionsschätzungen Ergebnisse auch für solche Bereiche zu gewinnen sind, für die direkte Erhebungen nicht vorliegen.

Page 107:

In diesen Fällen müssen Schattenpreise bestimmt werden, das heißt Preise, die potenziellen Marktpreisen nahe kommen. Hierbei gibt es in der Regel nicht nur einen einzigen Wert. Vielmehr sind (z.B. je nach unterstellter Nutzung) mehrere alternative fiktive Preise möglich, die auf mehr oder weniger zweifelhaften Annahmen beruhen können.

Page 107:

Ein wirklicher Marktwert existiert meist nicht, und damit ist man auf Schätzungen angewiesen, die aber in Wirklichkeit wegen ihrer prinzipiellen Unüberprüfbarkeit Festsetzungen sind. [...] Hier werden Zahlen eingesetzt, mit denen man nicht viel anfangen kann, und damit bekommen Begriffe wie Löhne, Gewinne usw. eine Bedeutung, die sie im normalen Sprachgebrauch nicht haben und die man nur noch nach eingehendem Studium der angewandten Berechnungsmethoden erfassen kann.

Page 114:

Wir haben ein Wirtschaftssystem vor uns, in dem sich die Wirtschafts- wie Unternehmerakte auf eine mehr oder minder genaue Kenntnis und eine in verschiedenem Grade gerechtfertigte Vermutung von dem Verhalten, das ist der Gesamtheit der ökonomischen Akte, der betreffenden Individuen stützen.

Page 114:

In diese Situation kommt nun die Totalprognose: es wird von einer Stelle, die volles Vertrauen genießt, eine Lagebestimmung des Gesamtsystems für eine an die gegenwärtige unmittelbar anschließende Zeitperiode vorgenommen. [...] Vorausgesagt wird also eine Gesamtheit von Wirtschafts- und Unternehmerakten. Diese Voraussage wird aber von den Trägern dieser zukünftigen Wirtschaftsakte als richtig angenommen. Es tritt also in jede einzelne Individualprognose ein neues Datum ein: Das Wissen um das Verhalten der anderen. [...] Das neue Datum ist unter allen Umständen so wichtig, daß es alles wirtschaftliche Verhalten berühren muß. [...] Die Wirtschafter und Unternehmer werden also [...] eine Umstellung ihrer Dispositionen — oder allgemeiner: Entscheidungen — vornehmen.

Page 115:

[Das] Ausmaß der wirklichen Veränderungen ist unter allen Umständen viel größer als das der vorausgesagten. Die Dauer der wirklichen Geschehnisse kann mit der vorausgesagten nicht zusammenfallen.

Page 115:

Man muß also die Frage erwägen, ob nicht eine solche Berichtigung der Prognose in Frage kommt, damit sie schließlich doch wahr werde. Wird die erste Prognose bekannt gegeben [...], so wäre die Möglichkeit vorhanden, dieser Prognose eine zweite [...] nachfolgen zu lassen. Die zweite Prognose würde das Vorhandensein der ersten als Datum für die Wirtschaft anerkennen und damit zu rechnen haben.

Page 179:

So gilt für [...] zielgerichtete Handlungen, dass sie um so rationaler sind, je besser der mit ihnen verknüpfte Anspruch auf [...] Effizienz begründet werden kann.

Page 205:

Durch das Hochschrauben der Wissenschaftlichkeitsstandards wird der Kreis anerkannter und damit handlungsrelevanter Risiken minimiert, und folglich werden implizit wissenschaftliche Freibriefe der Risikopotenzierung erteilt. Zugespitzt formuliert: Das Bestehen auf der 'Reinheit' der wissenschaftlichen Analyse führt zur Verschmutzung und Verseuchung von Luft, Nahrungsmitteln, Wasser und Boden, Pflanze, Tier und Mensch.

Page 218:

Entwicklungsvarianten können also die Zukunft verbauen oder offenlassen. Je nachdem wird also eine Entscheidung für oder gegen eine Reise ins unbekannte Niemandsland der zwar ungesehenen, aber absehbaren 'Nebenfolgen' getroffen. Wenn der Zug einmal abgefahren ist, ist es schwer, ihn wieder anzuhalten. Wir müssen also Entwicklungsvarianten wählen, die die Zukunft nicht verbauen und den Modernisierungsprozess selbst in einen Lernprozess verwandeln, in dem durch die Revidierbarkeit der Entscheidungen die Zurücknahme später erkannter Nebenwirkungen immer möglich bleibt.

Page 218:

Wir müssen also praktische Entwicklungen daraufhin untersuchen, ob sie einen Risiko-Gigantismus enthalten, der den Menschen seiner Menschlichkeit beraubt und ihn von nun an bis in alle Ewigkeit zur Fehlerfreiheit verdammt.

Page 239:

Es ist [...] nicht das Versagen einzelner Wissenschaftler oder Disziplinen, sondern liegt systematisch in dem institutionell-methodischem Zugriff der Wissenschaften auf Risiken begründet. Die Wissenschaften sind so, wie sie verfasst sind — in ihrer überspezialisierten Arbeitsteiligkeit, in ihrem Methoden- und Theorieverständnis, in ihrer fremdbestimmten Praxisabstinenz —, gar nicht in der Lage, auf die Zivilisationsrisiken angemessen zu reagieren [...].

Page 242:

Die Möglichkeit der Selbstkontrolle, die alle Besitzer von Monopolen hochhalten, müssen ergänzt werden durch die Möglichkeit der *Selbstkritik.* Das heißt: was bisher sich nur mühsam gegen die Dominanz von Professionen oder betrieblichem Management einen Weg freikämpfen kann, muss *institutionell abgesichert* werden: Gegenexpertise, alternative Berufspraxis, innerberufliche und -betriebliche Auseinandersetzungen um Risiken eigener Entwicklungen, verdrängter Skeptizismus. In diesem Fall hat Popper wirklich recht: Kritik bedeutet Fortschritt. Nur dort, wo Medizin gegen Medizin, Atomphysik gegen Atomphysik, Humangenetik gegen Humangenetik, Informationstechnik gegen Informationstechnik steht, kann nach außen hin übersehbar und beurteilbar werden, welche Zukunft hier in der Retorte ist. Die Ermöglichung von Selbstkritik in allen Formen ist nicht etwa eine Gefährdung, sondern der wahrlich einzige Weg, auf dem der Irrtum, der uns sonst früher oder noch früher die Welt um die Ohren fliegen lässt, vorweg entdeckt werden könnte.

Page 244:

Das heute anerkannte Wissen um die Risiken und Gefährdungen der wissenschaftlich-technischen Zivilisation hat sich überhaupt erst gegen die massiven Leugnungen, gegen den oft erbitterten Widerstand einer — selbstgenügsam und borniert in Fortschrittsgläubigkeit befangenen — 'wissenschaftlich-technischen Rationalität' durchgesetzt. Überall hinkt die wissenschaftliche Risikoforschung der sozialen Umwelt-, Fortschritts- und Kulturkritik am Industriesystem hinterher.

Page 247:

Der Rationalitätsanspruch, den Risikogehalt des Risikos sachlich zu ermitteln, entkräftet sich permanent selbst: [...] Zum anderen muss man einen Wertstandpunkt bezogen haben, um überhaupt sinnvoll über Risiko reden zu können.

Page 247:

In der Beschäftigung mit Risiken haben die Naturwissenschaften ungesehen und ungewollt sich selbst ein Stück *entmachtet, zur Demokratie gezwungen.*

Page 253:

[Diese Entwicklung] enthält die Chance der Emanzipation gesellschaftlicher Praxis *von* Wissenschaft *durch* Wissenschaft; andererseits *immunisiert* sie gesellschaftlich geltende Ideologien und Interessenansprüche gegen wissenschaftliche Aufklärungsansprüche und öffnet einer *Feudalisierung* wissenschaftlicher Erkenntnisansprüche durch ökonomisch-politische Interessen und 'neue Glaubensmächte' Tor und Tür.

# Bibliography


Angner, E. (2004), 'Revisiting Rawls: A Theory of Justice in the Light of Levi's Theory of Decision', *Theoria* 70(1), 3–21.

Anosov, O. L. & Butkoskii, O. Y. (1996), A Discriminant Procedure for the Solution of Inverse Problems for Nonstationary Systems, *in* Y. A. Kravtsov & J. B. Kadtke, eds, 'Predictability of Complex Dynamical Systems', Springer, pp. 3–19.

Armstrong, J. S. (1984), Forecasting with Econometric Methods: Folklore versus Fact, *in* S. G. Makridakis, ed., 'The Forecasting Accuracy of Major Time Series Models', John Wiley, chapter 2, pp. 19–34.

Baehr, H. D. (2000), *Thermodynamik*, Springer.

Bank of England (May 2003), *Inflation Report.*

Barrell, R. (2001), Forecasting the World Economy, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 9, pp. 19–44.

Beck, U. (1986), *Risikogesellschaft - Auf dem Weg in die Moderne*, Suhrkamp.

Beinhocker, E. D. & Kaplan, S. (2002), 'Tired of Strategic Planning?', *The McKinsey Quarterly. Special Edition Risk and Resilience* pp. 48–57.

Bestuzhev-Lada, I. (1993), Chapter 8, *in* Y. A. Kravtsov, ed., 'Limits of Predictability', Springer, pp. 205–222.

Betz, G. (2004), Apriorische und empirische Grenzen von Wirtschaftsprognose: Oskar Morgenstern nach 70 Jahren Prognoseerfahrung, *in* U. Frank, ed., 'Wissenschaftstheorie in Ökonomie und Wirtschaftsinformatik', DUV, pp. 171–190.

Blix, M., Wadefjord, J., Wienicke, U. & Adahl, M. (2001), 'How Good is the Forecasting Performance of Major Institutions?', *Sveriges Riksbank Economic Review* 2001(3), 38–68.

Britton, E., Fisher, P. & Whitley, J. (1998), 'The Inflation Report projections: understanding the fan chart', *Bank of England Quarterly Bulletin* 1998(February), 30–37.

Brümmerhoff, D. (2002), *Volkswirtschaftliche Gesamtrechnungen*, 7th edn, Oldenbourg.

Bryan, L. L. (2002), 'Just-in time strategy for a turbulent world', *The McKinsey Quarterly. Special Edition Risk and Resilience* pp. 16–27.

Bunsen, R. (1870), 'Calorimetrische Untersuchungen', *Poggendorfs Annalen* 141(1), 1–31.

Burns, T. (2001), The Cost of Forecast Errors, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 10, pp. 170–184.

Camba-Mendez, G., Kapetanios, G., Weale, M. R. & Smith, R. J. (2002), The Forecasting Performance of the OECD Composite Leading Indicators for France, Germany, Italy, and the UK, *in* M. P. Clements & D. F. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, chapter 17, pp. 386–408.

Campbell, J. Y., Lo, A. W. & MacKinlay, A. C. (1996), *The Econometrics of Financial Markets*, Princeton University Press.

Cao, L. & Soofi, A. S. (1999), 'Nonlinear deterministic forecasting of daily dollar exchange rates', *International Journal of Forecasting* 15, 421–430.

Cardwell, D. S. L. (1971), *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age*, Heinemann.

Carson, C. S. (1995), Design of Economic Accounts and the 1993 System of National Accounts, *in* J. W. Kendrick, ed., 'The New System of National Accounts', Kluwer, chapter 2, pp. 1–24.

Cartwright, N. (1983*a*), Do the Laws of Physics state the facts?, *in* 'How the Laws of Physics lie', Oxford University Press, chapter 3.

Cartwright, N. (1983*b*), *How the Laws of Physics lie*, Oxford University Press.

Cartwright, N. (1999), *The Dappled World: A Study of the Boundaries of Science*, Cambridge University Press.

Cawley, R., Hso, G.-H. & Salvino, L. W. (1996), Method to discriminate Against Determinism in Time Series Data, *in* Y. A. Kravtsov & J. B. Kadtke, eds, 'Predictability of Complex Dynamical Systems', Springer, pp. 3–19.

Chamberlin, E. H. (1948), 'An experimental imperfect market', *Journal of Political Economy* 56(2), 95–108.

Chermack, T. J. (2003), 'Improving Decision-Making with Scenario-Planning', *Futures*.

Chopard, B. & Droz, M. (1998), *Cellular Automata Modeling of Physical Systems*, Cambrige University Press.

Clarida, R. H., Sarno, L., Taylor, M. P. & Valente, G. (2001), 'The Out-of-sample Success of Term Structure Models as Exchange Rate Predictors: a Step Beyond', *NBER Working Paper, No. 8601*.

Clausius, R. (1850), 'Ueber die bewegende Kraft der Wärme und die Gesetze welche sich daraus für die Wärmelehre selbst ableiten lassen', *Annalen der Physik* 79, 368–397, 500–524.

Clements, M. P. & Hendry, D. F. (2002*a*), *A Companion to Economic Forecasting*, Blackwell.

Clements, M. P. & Hendry, D. F. (2002*b*), An overview of economic forecasting, *in* M. P. Clements & D. F. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, chapter 1, pp. 1–18.

Coyle, D. (2001), Making Sense of Published Economic Forecasts, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 4, pp. 19–44.

Crooks, E. & Major, T. (2003), 'Trichet's chance: can the European Central Bank's new chief improve communication and engineer recovery?', *Financial Times (6th November 2003)* p. 19.

Croushore, D. (1997), 'The Livingston Survey: Still Useful After All These Years', *Business Review, Federal Reserve Bank of Philadelphia* 1997(March/April), 1–12.

Darmstaedter, L. (1908), *Handbuch zur Geschichte der Naturwissenschaften und der Technik*, Springer.

Degrér, H., Hansen, J. & Sellin, P. (2001), 'Evaluation of exchange rate forecasts for the krona's nominal effective exchange rate', *Sveriges Riksbank Working Paper*.

DESTATIS (2002), 'Volkswirtschaftliche Gesamtrechnungen, Inlandsprodukt und Nationaleinkommen', *Reihe Methoden und Verfahren, Statistisches Bundesamt*.

Dicke, H. & Glismann, H. H. (2002*a*), 'Haben sich die Konjunkturprognosen des Sachverständigenrates verbessert?', *Wirtschaftsdienst Zeitschrift für Wirtschaftspolitik of the HWWA* 82(12), 736–740.

Dicke, H. & Glismann, H. H. (2002*b*), 'Konjunkturprognosen und wissenschaftlich-technischer Fortschritt', *Wirtschaftsdienst Zeitschrift für Wirtschaftspolitik of the HWWA* 82(3), 167–169.

Ditmars, D. A. (1984), Heat-Capacity Calimetry by Method of Mixtures, *in* 'Compendium of Thermophysical Property Measurement Methods', Plenum Press, chapter 13, pp. 527–553.

Essig, H. & Hartmann, N. (1999), 'Revision der Volkswirtschaftlichen Gesamtrechnungen 1991-1998', *Wirtschaft und Statistik* 99(6), 449–478.

European Commission (2000), *Communication of the Commission on the Precautionary Principle, COM(2000)1*.

European Environmental Agency (2001), *Late lessons from early warnings: the precautionary principle 1896-2000*, Office for Official Publications of the European Communities.

Faber, M., Manstetten, R. & Proops, J. (1996), *Ecological Economics: Concepts and Methods*, Edward Elgar.

Feyerabend, P. (1978), *Science in a Free Society*, NLB.

Fildes, R. & Ord, K. (2002), Forecasting Competitions: Their Role in Improving Forecasting Practice and Research, *in* M. P. Clements & D. F. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, chapter 15, pp. 322–353.

Fisher, I. (1925), 'Our unstable Dollar and the So-called Business Cycle', *Journal of the American Statistical Association* 20, 179–202.

Fixler, D. J. & Grimm, B. T. (2002), 'Reliability of GDP an Related NIPA Estimates', *Survey of Current Business* 2002(January), 9–27.

Ford, J. L. (1990), Shackle's Theory of Decision-Making under Uncertainty: A Brief Exposition and Critical Assessment, *in* S. F. Frowen, ed., 'Unknowledge and Choice in Economics', Macmillan, pp. 20–45.

Friedman, M. (1940), 'Review of J. Tinbergen: Statistical Testing of Business Cycle Theories, Vol. II: Business Cycles in the United States of America 1919-1932', *American Economic Review* 30, 657–660.

Frydman, R. & Phelps, E. (1983), Introduction, *in* R. Frydman & E. Phelps, eds, 'Individual Forecasting and aggregate outcomes', Cambridge University Press, pp. 1–30.

Funtowicz, S. O. & Ravetz, J. R. (1990), *Uncertainty and Quality in Science for Policy*, Kluwer.

Funtowicz, S. O. & Ravetz, J. R. (1993), 'Science for the Post-Normal Age', *Futures* 25(7), 739–755.

Funtowicz, S. O. & Ravetz, J. R. (1994), 'The Worth of a Songbird: Ecological Economics as a Post-normal Science', *Ecological Economics* 10(3), 197–207.

Gerhardt, M. & Schuster, H. (1995), *Das Digital Universum: Zelluläre Automaten als Modelle der Natur*, Vieweg.

Gillies, D. (2000), *Philosophical Theories of Probability*, Routledge.

Gillispie, C. C. (1971), *Dictionary of Scientific Bibliography*, Charles Scribner's Sons.

Gode, D. K. & Sunder, S. (1993), 'Allocative Efficiency of Markets with Zero-intelligence Traders: Market as a partial substitute to individual rationality', *Journal of Political Economy* 101, 119–137.

Gollier, C. & Treich, N. (2003), 'Decision-Making Under Scientific Uncertainty: The Economics of the Precautionary Principle', *The Journal of Risk and Uncertainty* 27(1), 77–103.

Granger, C. W. J. (2001), Evaluation of Forecasts, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 6, pp. 19–44.

Grossman, S. (1978), 'Further Results on the Informational Efficiency of Competitive Stock Markets', *Journal of Economic Theory* 18, 81–101.

Grunberg, E. & Modigliani, F. (1954), 'The Predictability of Social Events', *Journal of Political Economy* 62(6), 465–478.

Gyftopoulos, E. P. (1991), *Thermodynamics: Foundations and Applications*, Macmillan.

Habermas, J. (1981), *Theorie des kommunikativen Handelns, Band 1*, Suhrkamp.

Hall, R. E. (1978), 'Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence', *Journal of Political Economy* 86(6), 971–987.

Hamer, G. (1970), 'Genauigkeitskontrollen bei der Aufstellung volkswirtschaftlicher Gesamtrechnungen', *Allgemeines Statistisches Archiv* 54, 76–90.

Hansson, S. O. (1997), 'The Limits of Precaution', *Foundations of Science* 1997(2), 293–306.

Harsanyi, J. C. (1975), 'Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls' Theory', *American Political Science Review* 69, 594–606.

Healy, S. (1999), 'Extended peer communities and the ascendance of post-normal politics', *Futures* 31(7), 655–669.

Heller, A. (2001), 'Exploring the Fundamental Limits of Simulations', *Science and Technology Review*.

Hendry, D. F. (2001), How Economists Forecast, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 2, pp. 19–44.

Hendry, D. F. & Ericsson, N. R. (2001), *Understanding Economic Forecasts*, MIT Press.

Hobsbawm, E. J. (1977), *The Age of Revolution*, Sphere Books.

Holt, C. A. (1995), Industrial Organization: A Survey of Laboratory Research, *in* J. H. Kagel & A. E. Roth, eds, 'Handbook of Experimental Economics', Princton University Press, pp. 349–444.

Hong, J. T. & Plott, C. R. (1982), 'Rate Filing Policies for Inland Water Transportation: An Experimental Approach', *Bell Journal of Economics* 13, 1–19.

Hoyningen-Huene, P. (1998), Kuhn, Thomas Samuel, *in* E. Craig, ed., 'Routledge Encyclopedia of Philosophy', Routledge.

Huxley, A. (1947), *Science, Liberty and Peace*, Chatto & Windus.

IPCC (2001*a*), *Climate Change 2001: Mitigation; Contribution of the Working Group 3 to the Third Assessment Report of the IPCC*, Cambridge University Press.

IPCC (2001*b*), *Climate Change 2001: The Scientific Basis; Contribution of the Working Group 1 to the Third Assessment Report of the IPCC*, Cambridge University Press.

Jacobsen, C. & Bronson, R. (1997), Computer Simulated Empirical Tests of Social Theory: Lessons from 15 Years' Experience, *in* R. Conte, R. Hegselmann & P. Terno, eds, 'Simulating Social Phenomena', Springer, pp. 97–102.

Jaeger, C. C., Renn, O., Rosa, E. A. & Webler, T. (1998), Decision analysis and rational action, *in* S. Rayner & E. Malone, eds, 'Human Choice and Climate Change, Tools for Policy and Analysis, Vol. 3', Battelle, pp. 141–216.

Joutz, F. & Stekler, H. (2000), 'An Evaluation of the Predictions of the Federal Reserve', *International Journal of Forecasting* 16, 17–38.

Kacapyr, E. (1996), *Economic forecasting: the state of the art*, Sharpe.

Kadtke, J. B. & Kravtsov, Y. A. (1996), Introduction, *in* Y. A. Kravtsov & J. B. Kadtke, eds, 'Predictability of Complex Dynamical Systems', Springer, pp. 3–19.

Kagan, D. N. (1984), Adiabatic Calorimetry, *in* 'Compendium of Thermophysical Property Measurement Methods', Plenum Press, chapter 12, pp. 527–553.

Kahnemann, D., Slovic, P. & Tversky, A. (1982), *A Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kendrick, J. W. (1970), 'The Historical Development of National-Income Accounts', *History of Political Economy* 2, 284–315.

Kendrick, J. W. (1995), Introduction and Overview, *in* J. W. Kendrick, ed., 'The New System of National Accounts', Kluwer.

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan.

Knight, F. (1921), *Risk, uncertainty and profit*, Houghton Mifflin.

Kravtsov, Y. A. (1993), Chapter 7, *in* Y. A. Kravtsov, ed., 'Limits of Predictability', Springer, pp. 173–204.

Krelle, W. (1967), *Volkswirtschaftliche Gesamtrechnungen*, 2nd edn, Duncker & Humblot.

Krugman, P. (1998), How I work, *in* M. Szenberg, ed., 'Passion and Craft: Economists at Work', University of Michigan Press, pp. 143–154.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, 1996 edn, University of Chicago Press.

Laudan, L. (1981), 'A Confutation of Convergent Realism', *Philosophy of Science* 48, 19–48.

Lavoisier, A.-L. (1789), *Traité élémentaire de chimie*, Cuchet.

Ledyard, J. O. (1995), Public Goods: A Survey of Experimental Research, *in* J. H. Kagel & A. E. Roth, eds, 'Handbook of Experimental Economics', Princeton University Press, pp. 111–194.

Levi, I. (1980), *The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability and Chance*, MIT Press.

Levin, J. (1995), Government in the 1993 System of National Accounts, *in* J. W. Kendrick, ed., 'The New System of National Accounts', Kluwer, chapter 6, pp. 1–24.

Lorenz, H.-W. (1993), *Nonlinear Dynamical Economics and Chaotic Motion*, 2nd edn, Springer.

Loungani, P. (2001), 'How Accurate are Private Sector Forecasts? Cross-country Evidence from Consensus Forecasts of Output Growth', *International Journal of Forecasting* 17, 419–432.

Lovelock, J. E. (1987), *Gaia: A New Look at Life on Earth*, 2nd edn, Oxford University Press.

Lucas, R. E. (1976), 'Econometric Policy Evaluation: A Critique', *Carnegie-Rochester Conference Series on Public Policy* 1, 19–46.

Lycan, W. G. (2002), Explanation and Epistemology, *in* P. K. Moser, ed., 'The Oxford Handbook of Epistemology', Oxford University Press, chapter 14, pp. 408–433.

Makridakis, S. G. (1984), Forecasting: State of the Art, *in* S. G. Makridakis, ed., 'The Forecasting Accuracy of Major Time-series Methods', John Wiley, chapter 1, pp. 1–18.

Makridakis, S. G. & Hibon, M. (2000), 'The M3-Competition: Results, Conclusions and Implications', *International Journal of Forecasting* 16, 451–476.

Malkiel, B. G. (1992), Efficient Market Hypothesis, *in* 'The Palgrave Dictionary of Money and Finance', Palgrave Macmillan, pp. 739–742.

Malkiel, B. G. (1999), *A Random Walk down Wallstreet*, 7th edn, W.W. Norton.

Mann, M. E., Bradley, R. S. & Hughes, M. K. (2004), 'Global-scale Temperature Patterns and Climate Forcing over the Past Six Centuries', *Nature* 430, 105.

Marchi, B. D. & Ravetz, J. R. (1999), 'Risk Management and Governance: a Post-normal Science Approach', *Futures* 31(7), 743–757.

Mariano, R. S. (2002), Testing Forecast Accuracy, *in* M. P. Clements & D. F. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, chapter 13, pp. 284–298.

Mark, N. C. (1995), 'Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability', *American Economic Review* 85, 201–212.

Marshall, A. (1920), *Principles of Economics*, 8th edn, Macmillan.

McCracken, M. W. & West, K. D. (2002), Inference about Predictive Ability, *in* M. P. Clements & D. F. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, chapter 14, pp. 299–312.

McGuffie, K. & Henderson-Sellers, A. (2001), 'Forty Years of Numerical Climate Modelling', *International Journal of Climatology* 21, 1067–1109.

McIntyre, S. & McKitrick, R. (2003), 'Corrections to the Mann et. al. (1998) Proxy Data Base and Northern Hemisphere Average Temperature Series', *Energy and Environment* 14(6), 751–771.

McNees, S. K. (1992), 'How Large are Economic Forecast Errors?', *New England Economic Review* 1992(July/August), 25–42.

Médard, L. & Tachoire, H. (1994), *Histoire de la Thermochimie*, Université de Provence.

Meese, R. A. & Rogo, K. (1983*a*), 'Empirical Exchange Rate Model of the Seventies', *Journal of International Economics* 14, 3–24.

Meese, R. A. & Rogo, K. (1983*b*), The Out-of-sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification?, *in* J. A. Frenkel, ed., 'Exchange Rate and International Economics', University of Chicago Press.

Meese, R. A. & Rogoff, K. (1988), 'Was it real? The Exchange Rate-Interest Differential Relation over the Modern Floating Rate Period', *Journal of Finance* 43, 933–948.

Meyer, W. B., Butzer, K. W., Downing, T. E., II, B. T., Wenzel, G. W. & Wescoat, J. L. (1998), Reasoning by Analogy, *in* S. Rayner & E. Malone, eds, 'Human Choice and Climate Change, Tools for Policy and Analysis, Vol. 3', Battelle, pp. 217–290.

Middleton, C. (2002), 'Emerging from the Cold War – Stockpile Stewardship and Beyond', *Science and Technology Review* 2002(December), 4–13.

Morgan, M. S. (1990), *The History of Econometric Ideas*, Cambridge University Press.

Morgenstern, O. (1928), *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*, Springer.

Morris, J. (2000), Defining the Precautionary Principle, *in* 'Rethinking Risk and the Precautionary Principle', Butterworth Heinemann.

Muth, J. F. (1961), 'Rational Expectations and Theory of Price Movements', *Econometrica* 29(3), 315–335.

National Association for Business Economics (2003), *About NABE, http://www.nabe.com/members.htm, July 2004.*

Neubauer, W. (1994), Wirtschaftsstatistik und Konzepte der Volkswirtschaftlichen Gesamtrechnungen, *in* Statistisches Bundesamt, ed., 'Volkswirtschaftliche Gesamtrechnungen: bewährte Praxis und neue Perspektiven', Metzler-Poeschel.

Newbold, P. & Granger, C. W. (1974), 'Experience with Forecasting Univariate Time Series and the Combination of Forecasts', *Journal of Royal Statistical Society A* 137, 131–165.

Newhouse, S., Ruelle, D. & Takens, F. (1978), 'Occurance of Strange Axiom-A Attractors Near Quasiperiodic Flow on $T^n$', *Comm. Math. Phys.* 64, 35.

Nicolis, G. & Prigogine, I. (1998), *Exploring Complexity. An Introduction*, 5th edn, W. H. Freeman.

Ochs, J. (1995), Coordination Problems, *in* J. H. Kagel & A. E. Roth, eds, 'Handbook of Experimental Economics', Princeton University Press, pp. 195–252.

O'Connor, M. (1999), 'Dialogue and Debate in a Post-normal Practice of Science: a Reflexion', *Futures* 31(7), 671–687.

Öller, L.-E. & Barot, B. (2000), 'The Accuracy of European Growth and Inflation Forecasts', *International Journal of Forecasting* 16, 293–315.

Ormerod, P. (2000), *Butterfly Economics: A New General Theory of Social and Economic Behavior*, 2nd edn, Pantheon Books.

Osborn, D. R., Sensier, M. & Simpson, P. W. (2001), Forecasting and the UK Business Cycle, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 7, pp. 19–44.

Papineau, D., ed. (1996), *The Philosophy of Science*, Oxford University Press.

Pesaran, M. H. (1987), *The Limits to Rational Expectations*, Basil Blackwell.

Putnam, H. (1975), *Philosophical Papers*, Vol. 1, Cambridge University Press.

Putnam, H. (1981), *Reason, Truth and History*, Cambridge University Press.

Putnam, H. (2002), *The Collapse of the Fact/Value Dichotomy*, Harvard University Press.

Ravetz, J. (2004), 'The Post-normal Science of Precaution', *Futures* 36(3), 347–357.

Rawls, J. (1971), *A Theory of Justice*, Harvard University Press.

Rayner, S. & Malone, E. (1998), *Human Choice & Climate Change, Tools for Policy Analysis, Vol. 3*, Battelle Press.

Reid, D. J. (1969), A Comparative Study of Time Series Prediction Techniques on Economic Data, PhD thesis, Department of Mathematics, University of Nottingham.

Rescher, N. (1998), *Predicting the Future. An Introduction to the Theory of Forecasting*, State University of New York Press.

Richardson, M. J. (1984), Application of Differential Scanning Caliometry to the Measurement of Specific Heat, *in* 'Compendium of Thermophysical Property Measurement Methods', Plenum Press, chapter 17, pp. 527–553.

Romer, D. (1996), *Advanced Macroeconomics*, McGraw-Hill.

Rosenberg, A. (1995), *Philosophy of Social Science*, 2nd edn, Westview Press.

Rosenhead, J. (2001*a*), Robustness Analysis, *in* 'Rational Analysis for a Problematic World Revisited', 2nd edn, Wiley, pp. 181–207.

Rosenhead, J. (2001*b*), Robustness to the first degree, *in* 'Rational Analysis for a Problematic World Revisited', 2nd edn, Wiley, pp. 209–223.

Rosenhead, J. & Mingers, J. (2001), A New Paradigm of Analysis, *in* 'Rational Analysis for a Problematic World Revisited', 2nd edn, Wiley, pp. 1–20.

Roth, A. E. (1995), Introduction, *in* J. H. Kagel & A. E. Roth, eds, 'Handbook of Experimental Economics', Princeton University Press, pp. 3–109.

Salmon, W. (1989), *Four Decades of Scientific Explanation*, Vol. XIII of *Minnesota Studies in the Philosophy of Science*, University of Minnesota Press.

Samuelson, P. (1963), *Foundation of Economic Analysis*, Harvard University Press.

Schreiber, T. & Kantz, H. (1996), Observing and Predicting Chaotic Signals: Is 2% Noise Too Much?, *in* Y. A. Kravtsov & J. B. Kadtke, eds, 'Predictability of Complex Dynamical Systems', Springer, pp. 3–19.

Science & Environmental Health Network (1998), *The Wingspread Conference, http://www.sehn.org/wing.html, November 2003.*

Seydel, R. (1994), *Practical Bifurcation and Stability Analysis*, Springer.

Shackle, G. L. S. (1949), *Expectations in Economics*, Cambridge University Press.

Sheffrin, S. S. (1996), *Rational Expectations*, 2nd edn, Cambridge University Press.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, A. and C. Black.

Sonntag, R. E. & Wylen, G. J. V. (1991), *Introduction to Thermodynamics: classical and statistical*, John Wiley.

Stegmüller, W. (1973), *Personelle und Statistische Wahrscheinlichkeit. 'Jenseits von Popper und Carnap': Die logischen Grundlagen des statistischen Schließens*, Vol. IV, Part D of *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Studienausgabe*, Springer.

Stiglitz, J. E. (2002), *Globalization and Its Discontents*, Penguin.

Strohm, W. (1997), 'Beitrag der amtlichen Statistik zur gesamtwirtschaftlichen Konjunkturbeobachtung', *Wirtschaft und Statistik* 1997(10), 683–688.

Sullivan, R., Timmermann, A. & White, H. (1999), 'Data-Snooping, Technical Trading Rule Performance, And The Bootstrap', *Journal of Finance* 54, 1647–91.

The Conference Board (2003), *Business Cycle Indicators, http://www.globalindicators.org, July 2003.*

Turner, P. (2001), Economic Modeling for Fun and Profit, *in* D. F. Hendry & N. R. Ericsson, eds, 'Understanding Economic Forecasts', MIT Press, chapter 3, pp. 19–44.

United Nations (1992), *Report of the United Nations Conference on Environment and Development (A/CONF.151/26).*

U.S. Bureau of Economic Analysis (2002), 'Updated Summary NIPA Methodologies', *Survey of Current Business* 2002(October), 20–38.

van den Bogaard, A. (1999), Past Measurement and Future Prediction, *in* M. S. Morgan & M. Morrison, eds, 'Models as Mediators: Perspectives on Natural and Social Sciences', Cambridge University Press, chapter 10, pp. 282–325.

van der Waals, J. D. (1988), *On the Continuity of the Gaseous and Liquid State (reprint)*, North Holland.

von Hayek, F. A. (1972), *Die Theorie komplexer Phänomene*, J.C.B Mohr (Paul Siebeck).

Wagemann, E. (1927), *Kreislauf und Konjunktur der Wirtschaft*, Verlag Quelle u. Meyer.

Wagemann, E. (1928), *Konjunkturlehre - eine Grundlegung zur Lehre vom Rhythmus der Wirtschaft*, Hobbing.

Weidmann, J. (2002), 'Hat sich die Prognosetreffsicherheit des Sachverständigenrates systematisch verändert?', *Wirtschaftsdienst Zeitschrift für Wirtschaftspolitik of the HWWA* 82(12), 741–748.

Williams, B. (2002), *Truth and Truthfulness: an Essay in Genealogy*, Princeton University Press.

Wolfram, S. (1986), *Theory and Applications of Cellular Automata*, World Scientific.

Wynne, B. (1993), Uncertainty and Environmental Learning: Reconceiving Science and Policy in the Preventive Paradigm., *in* T. Jackson, ed., 'Clean Production Strategies: Developing Preventive Environmental Management in the Industrial Economy.', Lewis.

Young, A. H. (1995), Reliability and Accuracy of Quarterly GDP Estimates: A Review, *in* J. W. Kendrick, ed., 'The New System of National Accounts', Kluwer, chapter 13.

Zarnowitz, V. (1992), *Business Cycles: Theory, History, Indicators and Forecasting*, The University of Chicago Press.

Zarnowitz, V. & Braun, P. (1992), 'Twenty-two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance', *NBER Working Paper, No. 3965*.

# Index