International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

New Developments in the Methodology of Expert- and Argument-Based Probabilistic Population Forecasting

Lutz, W., Saariluoma, P., Sanderson, W.C. and Scherbov, S.

IIASA Interim Report
March 2000

IIASA
Science for
Global Insight

International Institute for Applied Systems Analysis
Schlossplatz 1 • A-2361 Laxenburg • Austria
Telephone: (+43 2236) 807 342 • Fax: (+43 2236) 71313
E-mail: publications@iiasa.ac.at • Internet: www.iiasa.ac.at

Interim Report **IR-00-020**

New Developments in the Methodology of Expert- and Argument-Based Probabilistic Population Forecasting

Wolfgang Lutz (lutz@iiasa.ac.at)
Pertti Saariluoma (psa@utu.fi)
Warren C. Sanderson (wsanderson@datalab2.sbs.sunysb.edu)
Sergei Scherbov (s.scherbov@frw.rug.nl)

Approved by

Gordon J. MacDonald (macdon@iiasa.ac.at)
Director

March 28, 2000

Contents

Abstract

All population projections are based in one form or another on expert judgement about likely future trends, structural continuity, etc. Although experts are clearly superior to lay people in their field of expertise, when it comes to forecasting they may make serious errors and can be as ignorant as anybody. In the context of expert-based probabilistic population projections, this issue is receiving even more attention. In this paper we argue that information about the future cannot be true for the reason that it is being presented by an acknowledged authority (institution or person), nor is it acceptable to resolve scientific issues on the ground of voting or concert. As an alternative we propose the concept of argument-based expert opinion. Under this approach any structural and trend assumption needs to be based on explicit argumentation rather than implicit judgement.

Acknowledgments

This paper was presented at the Population Association of America Annual Meeting, Los Angeles, California, USA, 23-25 March 2000.

About the Authors

Wolfgang Lutz is Leader of the Population Project at IIASA.

Pertti Saariluoma is Professor of Cognitive Science at the University of Helsinki, FIN-00014 University of Helsinki, Finland.

Warren C. Sanderson is Professor of Economics at the State University of New York, Stony Brook, NY 11794-4384, USA.

Sergei Scherbov is Professor of Demography at the Population Research Centre, University of Groningen, NL-9700 AV Groningen, the Netherlands.

The authors are experts in demography and cognitive science.

New Developments in the Methodology of Expert- and Argument-Based Probabilistic Population Forecasting

Wolfgang Lutz, Pertti Saariluoma, Warren C. Sanderson, Sergei Scherbov

1 Introduction

There is a large and continuing literature on the use of expert opinion in forecasting. The UN, the US Bureau of the Census, the International Institute for Applied Systems Analysis (IIASA), and most national statistical offices use expert opinion as a basis for their population forecasts. The World Bank uses a composite of estimated equations and expert opinion as inputs into their forecasts. Nevertheless, the forecasts and associated documents from these four international and most national organizations make no reference to the substantial literature on the use of expert opinion. Recently, an interesting debate has arisen in demography concerning the role of experts and expert knowledge in probabilistic population projections. Thus far, this debate has also ignored the larger literature. The thesis of this paper is that demographic forecasting should no longer stand alone. It has a great deal to learn from other forecasting experiences. In particular, the recent debate in demography about the use of experts in making probabilistic projections can best be understood in the context of the larger literature.

We set the stage in Section 2 by discussing why expert opinion has recently become an issue in population forecasting. A good summary of the current state of the art in forecasting is Armstrong (forthcoming). It contains 138 principles that cover all aspects of the forecasting process. In Section 3, we discuss twenty of those principles that pertain most closely to the use of experts in making probabilistic population forecasts. In Section 4, we present the foundation of a new approach to population forecasting that we call "expert- and argument-based probabilistic population forecasting." We are very far from having all the answers, but at this point we believe that even systematically investigating a formal forecasting framework for population is a step in the right direction. An understanding of the nature of expertise is essential for the proper application of expert opinion in forecasting. We consider this in Section 5. In Section 6, we discuss problems in getting knowledge from experts. With the new framework in mind, the recent debate on the use of expert opinion versus time series analysis in the production of prediction intervals is revisited in Section 7. Section 8 contains concluding remarks which point toward how the new approach could be implemented in practice.

2 Why Expert Judgment Has Recently Become an Issue in Population Forecasting

For decades most population forecasts made by national and international agencies have been based on the judgment of experts. This approach has been uncontroversial until recently, when an interesting debate has arisen concerning the role of experts and expert judgment in probabilistic population forecasting. Although the questions involved in this debate concern any kind of scientific forecasting–not only probabilistic and not only in the field of population–the competition of alternative approaches to probabilistic population forecasting has contributed to a sharper focus on some of the fundamental issues concerning the role of expert judgment.

A brief review of the debate is helpful in putting it into perspective. In Lutz, Sanderson and Scherbov (1997) we presented the first fully probabilistic projections of the world population by 12 major regions, which gives uncertainty intervals not only for population size but also for age distribution. It is based on thousands of simulations drawing from expert defined uncertainty distributions for fertility, mortality and migration in each of the regions. These uncertainty distributions had been derived from an interactive discussion process among a group of experts whose views and arguments are made explicit in a 500 page volume entitled *The Future Population of the World. What Can We Assume Today?* (Lutz 1996).

Tuljapurkar (1997) wrote a short commentary in which he supports the importance of probabilistic population projections but at the same time raises some doubts about the treatment of short term variations in vital rates and about the usefulness of expert opinion. These comments are based on some of the pioneering work of Lee and Tuljapurkar (1994) in developing a stochastic population projection model for the United States, which combines exogenous (expert) assumptions about the future trends in the demographic components with stochastic short term variations derived from the statistical analysis of past US time series drawing on earlier time series models on fertility (Lee 1993) and mortality (Lee and Carter 1992). Other roots of probabilistic population projections date back to Keyfitz (1985) and Stoto (1983), Pflaumer (1988) and Alho and Spencer (1985) suggesting alternative approaches based on ex-post error analysis, expert opinion and time series analysis, respectively.

An answer to some of the questions concerning the validity of expert opinion and the importance of the explicit modeling of short term fluctuations versus long term trends is given in Lutz and Scherbov (1998). They present various kinds of sensitivity analyses by empirically comparing the autoregressive (AR) processes advocated by time series proponents to a piece-wise linear probabilistic approach similar to the one suggested by Lutz, Sanderson and Scherbov (1997). This paper demonstrates that the model is rather robust with respect to minor modifications of expert views, and that the piece-wise linear probabilistic approach does not systematically underestimate the variance relative to the autoregressive method as has been suggested by Lee (personal correspondence). Lutz and Scherbov also discuss the practical usefulness and empirical feasibility, and come to the conclusion that the expert-based probabilistic approach is the only one applicable in settings with very limited time series data. It is structurally closer to the currently dominant method of projection variants, and has therefore the advantage of an evolutionary development.

The latest round of discussion on the issue of experts versus time series data is documented in a special supplement of *Population and Development Review* entitled

"Frontiers of Population Forecasting" (Lutz, Vaupel, Ahlburg 1999), with contributions by Lee and Lutz, Sanderson, Scherbov. A careful reading of both papers makes it clear that actually there are two independent questions that need to be separated for a better understanding and further progress in the field:

1. the choice of the process used in forecasting vital rates, autoregressive versus piece-wise linear.
2. the role of expert judgment in defining both the trend and the variance of the vital rates.

These are largely independent questions and it just happens to be the case due to intellectual traditions that authors who put more weight on expert judgment chose the piece-wise linear approach, while those relying more heavily on time series analysis chose the random walk as the preferred process. As shown by Lutz and Scherbov (1998) the choice of random walk as a process is also fully compatible with expert judgement as the basis for the assumed variance.

The choice of the process is (ironically) also a matter of expert judgement and to some degree a matter of taste. We agree with Lee that a random walk probably comes closer to the empirical nature of the real process, where we do observe irregular ups and downs. On the other hand, probabilistic scenarios may be a good and robust compromise between the current practice of completely disregarding probabilistic aspects and fully stochastic models because they are more in the tradition of current practice and seem to be easier to digest as a next step for the broad community of practitioners in statistical offices. The choice of the process only makes a minor difference for the outcome and is not the key issue.

The key issue, in the broadest sense, is whether experts can appropriately be used to provide the inputs needed for making probabilistic projections. There is a narrower issue as well. It is whether the methodology of eliciting expert opinion in Lutz, Sanderson, and Scherbov (1997) is the best available one, or whether there are improvements that can be made. Our positions on these issues are clear. Experts can appropriately be used to provide the inputs needed for making probabilistic projections, by using the argument-based approach discussed in Section 4. To our knowledge this methodology has never been used in the literature, not even in Lutz, Sanderson, and Scherbov (1997). It is the next evolutionary step in our research on expert-based probabilistic projections.

3 Learning From the Literature on Forecasting

Armstrong (forthcoming) summarizes what is known about forecasting in 138 principles.[1] In his section "Implementing Judgmental Methods," he gives the following advice: "In general, you need to ask the right people the right questions." Among the principles Armstrong lists in this section are: (1) "Pretest the questions you intend to use to elicit judgments forecasts," (2) "Frame questions in alternative ways," (3) "Ask experts to justify their forecasts in writing," and (4) "Obtain forecasts from heterogeneous experts." Combining forecasts is one of the best ways of improving forecast accuracy. Among the principles pertaining to the combination of forecasts,

[1] Each is numbered and comes with brief comments under the rubrics description, purpose, conditions, strength of evidence, and source of evidence.

Armstrong includes: (5) “Combine forecasts from approaches that differ,” (6) “Use many approaches (or forecasters), preferably at least five,” and (7) “Use formal procedures to combine forecasts.”

In his section on the evaluation of forecasting methods, Armstrong includes as principles: (8) “Describe potential biases by forecasters,” (9) “Provide full disclosure of methods,” and (10) “Compare forecasts made by different methods.” There is an important set of principles concerning the assessment of uncertainty. They include: (11) “Estimate prediction intervals,” (12) “Use objective procedures to assess uncertainty and to estimate explicit prediction intervals (PIs),” (13) “Write out reasons why the forecasts might be wrong,” (14) “Obtain good feedback about the accuracy of forecasts and the reasons why errors occurred,” (15) “Combine prediction intervals from alternative forecasting methods,” (16) “Use safety factors to adjust for overconfidence in the PIs, which are typically too narrow,” and (17) “Do not assess uncertainty in a traditional group meeting.” Principle (16) is based on evidence in Arkes (forthcoming) and Chatfield (forthcoming) that prediction intervals produced by both quantitative and judgmental techniques tend to be too narrow. Principle (12) is based on the possibility that judgmental prediction intervals might be the narrower of the two. Armstrong himself calls Principle (12) “speculative.” One simulation study concerned with the accuracy of population forecasts found that time series methods often produced prediction intervals that were considerably larger than the true ones (see Sanderson 1995). Finally, in his section on learning, Armstrong includes the following principles among others: (18) “Consider the use of adaptive forecasting models,” (19) “Seek feedback about forecasts,” and (20) “Establish a formal review process for forecasting methods.”

Very few of these 20 Principles are applied in population forecasting. A good population forecasting methodology does not necessarily agree with all those Principles, but it should seriously address the larger literature on forecasting. There is no good reason for demographic forecasters to remain so isolated from the larger forecasting community.

4 Towards a Methodology of Expert- and Argument-Based Probabilistic Population Forecasting

When information is elicited from experts and turned into probabilistic population forecasts, three groups of experts with different kinds of expertise are needed. We call the people in the relatively larger group, from which information is elicited, panel (or primary) experts. We call the people in the relatively smaller group, who receive the information and process it to form prediction intervals, implementation (or secondary) experts. Design (or tertiary) experts are those who design and supervise the process by which information flows between the two groups.

The job of the design expert is a difficult one. Using experts in forecasting is in many respects meta-scientifically problematic. There are a few obvious argumentation fallacies, which must be avoided. Unfortunately, sufficient attention has not always been paid to these phenomena, so that in classic Delphi, for example, one can sometimes find argumentatively non-optimal practices (Hahn *et al*. 1999; Linstone and Turoff 1975; Martino 1983; Parenté and Anderson Parenté 1987) that lead to serious biases and errors and thus decrease the quality of expert-based predictions.

The first argumentative fallacy hidden in Delphi-methods is the tacit acceptance of an authority argument. Experts normally have substantial authority and this is the reason they are used in forecasting. It is easy to think that any piece of knowledge presented by an expert should be incorporated into argumentation because it has been presented by an expert. However, no piece of knowledge can be true, because it is presented by an acknowledged authority (Naylor 1997). If authority argument would be acceptable, Aristotle's theory of the heart as a seat of mind would presumably still be valid. The acceptance of expert opinions as arguments neglects the limits of expertise. Experts may easily make errors. Therefore, in forecasting, it is important that any piece of the expert's knowledge is assessed on the basis of the arguments that he or she gives to support the presented view, and not on the basis of the presenter's personal expertise. The contents and power of arguments decide if the piece of knowledge is accepted into a model of prediction.

The second problem arises when several experts have different views. In this case, it is easy to think that voting or some other group decision process provides a sufficiently reliable basis for forecasting. Again, it is unacceptable to resolve scientific issues on the basis of voting or other group decision process. The majority argument is an equally elementary fallacy in argumentation as the authority argument. Truth cannot be determined by majority vote. It must be resolved on the basis of the arguments. Therefore, in principle, all voting-based methods of forecasting are wrong. If we had accepted scientific knowledge based on voting, we would never have changed our views about the solar system. Practically all new scientific ideas are in some sense against the current. Science is a system of truths, and new truths do not belong to the system of old truths, before they have been incorporated into it. In fact ever since Kuhn's (1962) famous model of scientific advancement, it has been generally known that one must make a distinction between anomalies and accepted truths. Anomalies are findings which do not fit together with the prevailing theory. Yet anomalies can entail important pieces of truths, but still be outside the generally accepted theoretical thinking. Consequently, anomalies will easily have an effect on voting-type predictions in a very late stage.

The problem with majority and authority arguments is that they make self-correction and falsification impossible, and without constant self-correction, there is no genuine science. Unless one is able to test the arguments, one cannot have any mechanism of self-correction (Popper 1959). However, to test an expert's knowledge one must know the presented arguments behind the beliefs. In forecasting, the actual events will emerge much later than the models and this is why arguments are so important to their testing. It is rarely possible to falsify a forecast without falsifying the arguments, because the falsification by showing the outcome incorrectly is often too late.

Let's assume that an expert gives a value of 1.3 for German TFR in 2025. How can we falsify this knowledge as an argument, if we think that the reason and justification for its correctness is that it has been presented by an expert? The basic fact does not change whatever happens. An expert has put it forward and it stays, unless she herself corrects the fact. So, if he or she remains convinced of the value whatever happens, the argument does not change. Of course, it is not rare that experts do not change their opinions. Many of Darwin's opponents, for example, never changed their minds, but after their death younger generation scientists accepted Darwin's views.

It is one thing to take an expert's estimate when there is some scientific argumentation for it, and another to accept it, because it has been put forward by an expert. Accepting that TFR will be 1.3 because of contraceptive improvements is a valid argument. Accepting the value because an expert said so is not science. You can always test the first fact and discuss it. You have no possibility to discuss the value, if you accept the expert's opinion as a valid argument. This is apparently a small difference, but in fact, it is essential. Opinions are simply not arguments, because one cannot possibly test them.

Expert opinions as arguments are problematic for several other reasons. The bases of expert opinions are not public. However, in science all the factors that affect the outcome of argumentation must be visible to anybody, so that this anybody can critically think them. Publicity is an absolute demand for science, because without it no self-correction is possible (Popper 1959). Experts may always have conscious or unconscious reasons for their opinions. They are people and they have political, cultural or personal values. This may affect their estimates. Unless, the arguments are absolutely public for all, it is impossible for the scientific community to test or critically analyze the arguments behind forecasts.

The core of all problems with some earlier practices of using experts in forecasting is their inherent subjectivity. They, in one or another way, incorporate expert opinions into their web of arguments and miss a subtle conceptual difference between expert opinions and knowledge. Expert opinions need not be argued, but their knowledge must be. Knowledge is objective only if its truth does not depend on the will of any human being. A proposition P is not necessarily true just because an expert believes it, even in the case when P is really true. If I believe that my lottery coupon will win and I indeed win, I still cannot say that I knew that the ticket would win. I just made a lucky guess. Lucky guesses are not knowledge, because they are not justified by acceptable arguments. From a scientific point of view, the lucky guesses are not sufficient, but arguments and justification via arguments are decisive. Science is thus knowledge about truths, reasons and arguments. This means that only arguments can make expert knowledge something more than simply lucky guesses.

To avoid the problems associated with the incorrect use of expert knowledge, the design experts must find a new type of elicitation mechanism based on expert knowledge instead of expert opinion. The latter are argumentatively all too weak to safeguard forecasts against biases and errors. Even the best can err, and if there is no objective mechanism for securing expert knowledge, the forecasting models made following expert opinions cannot be effectively tested. For these reasons a two-stage procedure for the use of expert knowledge will be suggested here. The main idea is to allow the incorporation of argumentation into expert-based forecasting.

In this model there are two basic stages. In the first stage, which can be called constructing, the design and implementation experts collect all possible views from the panel experts with the arguments supporting the opinions. In the second stage, termed proving, the arguments are analyzed by the implementation experts; the relevance of knowledge and arguments are discussed; and finally, the knowledge is incorporated into the theory or model, if the arguments are valid. This two-stage model should ensure that the argumentation is theoretically valid, but also give a much wider and accurate base of information for modeling, compared to the intuitions of the experts alone. The main thing is that the publicity and openness of the procedure makes scientific self-correction

possible. Demography deals with all too important issues to allow subjectivism in the form of non-openly argued frameworks.

The suggested way of using expert knowledge in forecasting means first, that panel experts are used to collect as many rational points of view as possible for the use of forecasting. Second, expertise as such is not used as an argument, but the arguments experts present are put under critical scrutiny and incorporated into the model and used in forecasting when it makes sense.

The three groups of experts (panel, implementation, and design experts) play very different roles and need very different kinds of expertise. The panel experts are needed for arguments about the likely demographic future. They should include people from a wide variety of fields, as long as they bring informed arguments about the future developments and how they could affect population change. If, for example, there were a significant probability that a massive meteor would collide with the Earth within the forecast period, astronomers should certainly be among the group of panel experts. Implementation experts need to be able to evaluate the arguments proposed by the panel experts and turn them into probabilistic forecasts. They do not need to include people with a deep understanding of every field. Implementation experts can call upon other experts to aid them in evaluating arguments. Because they need to be able to take arguments as inputs and turn them into demographic outputs, implementation experts are likely to come from demography, sociology, economics, and other related disciplines. Design experts are likely to come from disciplines such as cognitive psychology, forecasting, and sociology, guided, of course, by demographers. Expert- and argument-based forecasting is an inherently interdisciplinary business.

5 What is Expertise?

The discussion about the use of expert knowledge and its use in demography has seldom paid attention to the nature of expertise. However, a clear analysis of expertise is very helpful in constructing appropriate procedures for using experts in forecasting. In any domain, all the people are not equally expert, but the level of expertise varies (Chase and Simon 1973; de Groot 1965; Ericsson and Delaney 1998; Ericsson and Lehman 1996). There are high-level experts such as grand masters in chess, and less skilled like masters or novices. Normally, the highest level of skills are achieved after ten years of practice, and even child prodigies have worked this long in the field before reaching the highest levels of expertise. In the literature, this phenomenon, which has so systematically been observed, has been termed the “ten-years-rule” (Hayes 1981; Ericsson and Lehman 1996).

The measurable changes in an expert’s performance are very stable and systematic. Experts are much faster in attentional and motor tasks, and they normally discriminate targets much more accurately (Ericsson and Kintsch 1995; Saariluoma 1985). However, their general attentional capacity need not be any different from novices. In addition to perceptual and motor tasks, experts are superior in recall and recognition tasks, and also problem-solving tasks (Chase and Simon 1973; Ericsson and Kintsch 1995; Ericsson and Lehman 1996; de Groot 1965; Saariluoma 1985, 1995). Expertise is also domain specific, which means that a person can be an expert in some issues, but a totally normal person in all other respects (Chase and Simon 1973; Ericsson and Delaney 1999; Ericsson and Kintsch 1995; Ericsson and Lehman 1996; Vicente 1988; Vicente and Wang 1998).

Expertise does not keep people from making errors. Experts may err less frequently than novices, but they still err (Saariluoma 1992, 1995). In the literature, it is very easy to find cases in which experts have made serious errors in forecasting. A specialist predicted that "when the Paris world exhibition closes, the electric light will close with it, and no more will be heard of it." The chairman of the board of a large business manufacturer predicted in 1943: "I think there is a world market for about five computers" (Cerf and Navasky 1984). Given our goal of improving forecasting, it is good to understand why experts err.

There are many explanatory factors for an expert's errors. One is that their basic capacity of attention and memory is the same as that of a novice. Tasks may simply be all too complex and surpass their limited capacity, causing various types of errors (Johnson-Laird 1983; Reason 1990). Another explanation for an expert's errors is that their basic thought processes are not so different from the ones of novices. An expert's highest level thought structures are the same as those of novices. This means that experts can be biased by normal non-task specific illusions (Armstrong 1979). It is also quite possible that experts have wrong presuppositions; these presuppositions lead them to incorrect and ineffective problem representation (Saariluoma 1992, 1995). In short, experts can easily surpass the limits of their expertise. The risk of such problems is naturally largest with unknown future issues. The overconfidence that experts sometimes exhibit in their own opinions might be an example of this. For some experts, forecasting future demographic variables might be within their domain of expertise, but estimating the uncertainty of their forecasts might lie outside it.

Although the idea of using expert knowledge in forecasting is clear in theory, there are many practical problems in constructing expert-based forecasts, which one cannot leave outside the discussion. Even though the ultimate proof of argumentation is independent of the experts who have provided the knowledge, it is for practical reasons very important that experts really provide as much of their knowledge as possible.

To foresee problems and to think of possible solutions for them, it is important to use psychological methods and knowledge about experts. This is new ground. There is no direct knowledge about the practical problems associated with constructing argumentative models. Nevertheless, from the experiences from earlier expert-based models for forecasting and the information we have about constructing expert systems, for example, it is possible to foresee some of the practical issues that must be resolved. Some of the problems are associated with defining expertise and getting knowledge from experts. The rest are social psychological, and their origin is in group processes.

It is perhaps the most logical to begin with the practical issues arising from the determination of expertise. We do not have much knowledge about this, but in the psychology of computing, it has been considered in some depth. When working with expert knowledge, one has to decide who is an expert, in which issues he or she can be an expert, and what is the level of expertise. Without being careful in defining the type and the real competence of experts, it is easy to make illusory attributions. Even experts themselves may incorrectly define their competence areas, because the borders of competence are very difficult to define objectively in ill-defined domains.

One source of problems in measuring expertise is that expertise is domain specific and domains are normally wide. This means that a person can be a specialist in some sub-domain, but much less experienced in a sub-domain just next to the area of his or her main expertise. An expert on migration between the southwestern USA and Mexico

might be ignorant of the details of migration within the countries of Southern Africa. With respect to the likely future path of mortality, both may be as ignorant as a lay person. For these kind of reasons, it is very important to pay attention to the specifications of expertise.

Defining expertise is very easy in some task environments. Chess players and other sports people have relatively well-developed rating systems (Elo 1978). The main principle of the rating systems is quite clear: The stronger the opponent one is able to beat, the more the rating of a player is increased, and the weaker the player one loses to, the more points he or she will have deducted. When the outcome is as objective as it is in sports, the determination of expertise on the basis of rating is very efficient. Unfortunately, most of the professional domains on the limits of expertise may be very unclear.

Sometimes useful single measures of expertise cannot be found and one has to use several criteria. The time in the field, productivity, academic performances, and peer assessments can, for example, be used as the basis of evaluation. However, with all of these criteria it is impossible to reach such intermediate scale accuracy, which is typical of well-defined domains like chess. The time in the field is not an absolute measure, because experts with the same background may substantially vary in their competence. When comparing computer programmers with the same schooling, substantial differences have been found in their efficiency to produce code (Brooks 1980). Another problem is that people who have been in the field a long time have not necessarily done the required tasks, but have rather been in administrative tasks. Yet the time in the field is reasonable criterion, if it is used in a sensible manner and the possibly inappropriate preconditions have been taken into account. The differences between medium-level and high-level experts are normally rather large, and especially the differences between experts and novices are very striking.

A rough way of defining experts is academic schooling. It is normally better to take people who have academic degrees in the field than outsiders. However, academic fields are large and people specialized. It is not clear that an academic background in demography makes somebody a specialist, unless that person has really worked in the field. On the other hand, a sociologist or statistician who has really been working a long time with demographic issues may easily have the competence of an expert.

Productivity might also be a criterion. Even if people have relatively similar backgrounds, more productive people might have gained a better level of expertise because their motivation is higher. However, even here many sources of errors can be found. If one works with non-standard or otherwise more difficult problems, the quantitative productivity is a senseless way of assessing competence or motivation. Nevertheless, the sensible criteria for collecting a panel of experts should not be difficult to meet. One must just be cautious of possible loopholes in the selection process.

One way of approaching the expert panel selection problem is to use peer review. This means that one asks other demographers who they consider good experts. However, in computing research, it has been noticed that this method is subjective (Brooks 1980). The differences in schools, in the quality of personal relations, lapses of memory and so on may bias judgments. Nevertheless, normally the difference between high-level experts and medium-level experts can be made in this way.

In summary, there are several sources of biases in selecting experts for expert panels, but mostly they need not be devastating. By using multiple criteria, and by taking into account the possible sources of errors, reasonably effective selection procedures can be developed. The most important mechanism safeguarding against errors is, nevertheless, the use of expert knowledge in argumentation. Because the two-stage model does not take expert knowledge at face value, but builds on arguments, small errors in the selection of panel members are not so important. It is the weight of the arguments which decide, and not the level of expertise. One or two medium-level experts cannot change the final outcome, if their arguments are not strong. On the other hand, if they present strong arguments, it does not matter that they are not super experts. The avoidance of subjectivity is one of the major benefits of this model.

6 Obtaining Information From Experts

The goal of collecting expert knowledge is to get as wide and versatile a basis of information for forecasts as possible. Therefore, each of the used methodologies should have open elements. It is also important that researchers own *a priori* conceptions that do not prevent experts from freely bringing all their points of view, however imaginary, to the game. Consequently, it is very important to have in the research process elements which encourage experts to express all imaginable points of view. Possibly some derivative of brainstorming could be used in the beginning. Developing such a methodology requires much hard “grass roots level” work. We briefly discuss three possible approaches, the use of a panel, interviewing, and surveys. Our ideas here are tentative and suggestive.

One problem with this kind of panel may be to get people who are willing to commit themselves to this approach. In Delphi one tries to resolve these problems by using anonymous experts. The number of such specialists need not be large. The main thing is that they are willing to commit themselves to the process, and that they have a sufficiently relevant background with respect to the issues that are discussed. Indeed, it should be possible to look for the most efficient group working techniques.

The major goal of an expert panel is to get as good a view about the relevant information as possible. One can use several sessions, part of which should be innovative and brainstorming, and part of which can be more analytical. After the innovative stages, all presented points could be systematically discussed in groups. If differences of opinions appear, their backgrounds should be investigated as effectively as possible. Any antinomy is always a source of possibilities.

It is quite clear that it is difficult to get all the information needed in group sessions. Therefore, it might be wise to have interviews and even in-depth-interviews. The goal of these interviews would be to analyze critical issues and the basic intuitions of the experts. It is important to plan such interviews carefully. They must, on the one hand, give information about the well-known but problematic points. For example, if a particular expert sees something important, when others do not, it is important to understand why. However, it is also important to have openness, so that new types of factors can be collected. One must also be able to plan hypotheses about the reasons people have for their thinking, so that one can directly ask questions concerning the matter.

It is well known that surveys are good, because they allow the collection of information from a relatively large group of people at a reasonable cost. But it is not

possible to obtain equally in-depth knowledge with surveys as with interviews and panel techniques. One can use open questions and other respective ideas to improve their openness. Nevertheless, if the turnover is reasonable, surveys give a good idea how the experts think about the important issues in the field.

There are a number of techniques one should take into account concerning how to organize this kind of research. How to create a suitable atmosphere? How to arrange questions, what kind of questions should be asked, etc. These details can be left to further discussion. One important thing to remember: The purpose of this kind of research is not to get the distribution of ideas in the expert community, but only to get as many important points of view to be incorporated into the discussion as possible. This means that issues like sample biases are not important. The only important thing is to get a good base of arguments for making a model for rational forecasts.

There are some specific problems to be discussed concerning the nature of data. One can collect from the experts either numeric information or qualitative arguments. In accessing numeric information, a problem arises because of discrimination thresholds. People are not normally very accurate with quantitative information. People do not necessarily take a variation of around 10% very seriously, and they can thus make substantial errors in judging uncertain sums.

Another problem is that people do not normally remember quantities very precisely. Confusion is common, especially when the number of digits increases. Thus, it is easier to remember that a budget is 3.4 million rather than 3,486,196, for example (Katona 1940). One would imagine that experts are bit better in these kinds of issues, but experts may quite readily make errors. It is the task of methodology to decrease the risk of erring. Therefore, it would be good to ask experts to give clear reasons, why they think that some numeric value will be what they assess it to be. Of course, it is important that experts are given sufficient background information.

Experts should also evaluate how certain they are about the numbers they give. It may be that they feel forced to give an answer, although they are very uncertain and are practically guessing. However, they may also have very good grounds to support their judgments. In our methodology, uncertainty ranges must also be supported by argumentation. It may be the case that good arguments about uncertainty are not produced by the experts in the first round of elicitation. Arguments about uncertainty may become better articulated after the experts have had some feedback from earlier rounds, and see the range of ideas put forward by other experts. The main thing is that reasonable arguments be made concerning the uncertainty of used knowledge.

It is evident that a large part of an expert's contributions must be qualitative by nature. Though it is very commonly assumed that quantitative information is hard and qualitative "soft," the difference is not absolute. If a forecaster presents a numeric parameter to an expert for estimation, he or she has already had to decide that this parameter is important, and that the selection of parameters naturally presupposes qualitative analysis before quantification. The expert is not given the possibility to affect the choice of factors. On the other hand, if this is done, the qualitative nature of the data is obvious. Also, the information about possible phenomena and the mechanisms associated with these phenomena are information that must be classified as qualitative.

Qualitative information is thus a special position when methods of population projections are planned. When the researched phenomena are complex, qualitative

information is exceptionally valuable, because the selection of essential aspects is a qualitative decision. All factors incorporated into a model of forecasting belong to the class of qualitative knowledge. Even causal analysis presupposes qualitative decisions. Finally, the reasons asked from experts are normally qualitative by nature. Therefore, one must pay specific attention to the methods of qualitative information gathering with experts.

The methods of collecting qualitative information are normally quite explicit and today they are well known (see collection by Denzin and Lincoln 1998). One problem is that there is relatively little experience about the specific problems of applying qualitative methods, such as interviews and surveys with experts for solving problems like forecasting. We need to get more information about the specific characteristics of expertise in this kind of research situation. Nevertheless, some main principles can be discussed, even though the literature is relatively small.

Qualitative methodology provides the possibility to investigate issues that cannot normally be reached by quantitative methods, particularly reasons and mechanisms. Qualitative research methods have been designed to give information about mental and cultural contents, beliefs, opinions, interpretations, and intuitions. All of these are factors that may underlie an expert's assessment of factors, influencing their judgment in some context.

It is well known that qualitative data very often provides understanding rather than explanation. It gives insight into the heart of the matter rather than a clear explanation why things are as they are. It may simply be that people do not have the explanation (Hamilton 1994, 1998). Nevertheless, qualitative analyses of data are very suitable and effective, when the main goal of expertise research is to get the information about the arguments behind the opinions. These presuppositions are often tacit and intuitive (Saariluoma 1997). This means that one can best uncover them by means of deep interviews and group processes.

It is common to see qualitative and quantitative methods as some kind of mutual contradiction, but this is not necessarily the correct way to understand what they measure. No quantification is possible without first making a qualitative interpretation and then assigning a numeric value to the variation of the observed phenomena. Even measuring length presupposes qualitative analysis. Consequently, qualitative analysis and investigation, and collecting qualitative information about relevant demographic factors is the foundation of quantification. However, in forecasting it is often very important to turn qualitative knowledge collected from experts into quantitative measures and, therefore, qualitative analysis is not sufficient.

7 The Debate Revisited

With this background we can now revisit the recent debate about the use of experts in making probabilistic projections. The literature has shown that experts can be overconfident in the sense that prediction intervals can be too small. There are many possible reasons for this. It could be that experts have less expertise in forecasting prediction intervals than they have in forecasting most likely outcomes. The argumentation-based methodology that we propose does not simply ask experts how uncertain they are, it requires written (or recorded oral) arguments about the extent of their uncertainty.

Argumentation about uncertainty does not have to be elicited in the first round of information gathering. It seems likely to us that better arguments concerning uncertainty will appear over subsequent rounds after the experts have seen the arguments made by the entire group. For example, if strong reasons can be given to defend the position that life expectancy at birth in some region would be 80 years in 2050, and equally strong reasons can be given supporting the proposition that it would be 90 years, then the experts in the group will form their ideas about prediction interval for life expectancy at birth accordingly.

The group of panel experts includes people with a wide variety of expertise. It should include people who are more quantitatively oriented, such as time series analysts and structural modelers, as well as people who are primarily qualitatively oriented, and many people in between. The group members will make arguments for their points of view. Time series analysts and structural models could provide estimated equations to bolster their points of view. Sociologists could provide arguments concerning prediction intervals based on diffusion theory. Demographers might discuss, for example, uncertainty in fertility due to changes in the mean age at childbearing. Environmentalists might discuss uncertainties due to limitations in water supplies. Ethicists could bring up uncertainties due to the cloning of humans. When arguments matter (not just one's own faith in his/her ability to forecast) there is no particular reason to expect that experts would produce prediction intervals that are systematically too small or too large.

In order to make our ideas more concrete, it is useful to discuss a real example. Keilman and Hetlans (1999) used a particular time series model to estimate the 95 percent prediction interval for the total fertility rate (roughly the number of children ever born to the average woman over her lifetime) of Norway in 2050. That interval lay between 0.5 and 2.8 children. We cannot accept the Keilman and Hetlans estimate of the prediction interval for the reason that Keilman and Hetlans are experts. Nor can we accept their estimate because they used a quantitative method. Their argument, in fact, has many parts that we must evaluate. It is based primarily on the assumption that their particular model provides a very good representation of what happened during the period over which the parameters were estimated, and that it will provide an equally good representation onwards to the year 2050.

There can be two types of arguments made about their prediction interval, internal and external. An internal argument deals with their assumptions. For example, someone might argue that there are many plausible specifications that give very different results. In this case, Keilman and Hetlans would need to make an argument concerning why the specification that they chose was the superior one. Indeed the sensitivity of prediction intervals derived from time series analysis to the exact specification of the model is well known. Sanderson (1995), for example, showed this in the context of a demographic simulation experiment.

An external argument would approach the prediction interval from an alternative perspective. For example, a total fertility rate of 0.5 can be achieved by half the women in Norway having a single child, while the remaining half remain childless. This would certainly go well beyond the range of historical observation. There has never been a place or time where we have seen a situation in which there were no second or higher order births, not even in urban China, where their one-child family policy is most strictly enforced. If we would allow half of the women who had a first birth to have a second one (but no third or higher order births), then a total fertility rate of 0.5 could

only be observed if two-thirds of all Norwegian women were childless. Can a plausible argument be made that around two-thirds of all Norwegian women would be childless by 2050?

De Beer and Alders (1999) used a parity-based approach to derive a 95 percent prediction interval for the Netherlands in 2050. Their range is between 1.1 and 2.3 children, which is considerably smaller than the prediction interval for Norway produced in Keilman and Hetlans (1999). Is the fertility of Norwegian women really that much more uncertain than the fertility of Dutch women? If not, why is Keilman and Hetlans' prediction interval so much larger? These are the types of questions that get resolved through argumentation. No argument can be accepted because of who puts it forward or because of its methodology. All arguments need to be discussed and, if possible, tested. The important point that we would like to stress again is that there is no particular reason why a prediction interval based on arguments should be biased.

8 Concluding Comments

The skeletal methodology that we have presented here is very different from the standard use of expert opinion in forecasting. We propose a new approach based on arguments, not opinions. These arguments are derived from a group of panel experts who have knowledge relevant for the study of population change. Some of them might be time series modelers, and some might be practitioners of the Delphi method. Armstrong (forthcoming) suggests that questions be pre-tested and framed in alternative ways. These techniques will certainly be used in our framework. In the context of our approach, he recommends that experts report on their arguments in writing. We also believe that this is a good idea. We also believe that it is important to obtain arguments from a heterogeneous group of experts.

The essence of our approach is the combination of arguments made by various experts. It could be that arguments are combined at the level of the inputs into the forecasts (e.g. total fertility rates, life expectancies at birth, and migration rates) or that the forecasts based on a variety of arguments are subsequently combined. Our methodology is open. Full disclosure is one of its fundamental elements. A signature element of our methodology is that it produces prediction intervals. The overconfidence of experts is not a problem here because prediction intervals are based on arguments, not on subjective feelings. Prediction intervals are not produced in meetings where it is possible for some people to bully others, but through an impersonal approach based on anonymous arguments. The exact process of generating and evaluating such arguments needs explicit attention as a research project in its own right. Finally, arguments can be made stronger or weaker with additional evidence. Our argument-based approach is inherently adaptive and allows us to learn how to improve our forecasts over time. In these ways, the new expert- and argument-based methodology is more consistent with the literature on forecasting than anything currently being done.

Our next step is to perform a pilot projection study with a small group of volunteers. In dealing with the forecasting of complex phenomena like population change, there will never be a time when human expertise is not useful. But if we can replace the emphasis on expert opinion with one based on expert knowledge, we believe that we will taken a significant step toward improving our forecasts.

9 References

Alho, J.M. and B.D. Spencer. 1985. Uncertain population forecasting. *Journal of the American Statistical Association* 80(390):306-314.

Arkes, H. Forthcoming. Overconfidence in judgmental forecasting. In J.S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.
Currently available at http:\\www-marketing.wharton.upenn.edu\forecast\standard.pdf.

Armstrong, S.C. 1979. *Long-Range Forecasting. From Crystal Ball to Computer*. New York: Wiley.

Armstrong, J.S. Forthcoming. Standards and practices for forecasting. In J.S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.
Currently available at http:\\www-marketing.wharton.upenn.edu\forecast\standard.pdf.

Brooks, R. 1980. Studying programmer behaviour experimentally: The problem of proper methodology. *Communications of the ACM* 23:207-213.

Cerf, C. and V. Navasky. 1984. *The Experts Speak: The Definite Compendium of Authoritative Misinformation*. New York: Pantheon Books.

Chase, W.G. and H.A. Simon. 1973. The mind's eye in chess. Pages 215-281 in W. Chase (ed.), *Visual Information Processing*. New York: Academic Press.

Chatfield, C. Forthcoming. Prediction intervals for time series. In J.S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.
Currently available at http:\\www-marketing.wharton.upenn.edu\forecast\standard.pdf.

De Beer, J. and M. Alders. 1999. Probabilistic population and household forecasts for the Netherlands. Paper presented at the Joint ECE-Eurostat Work Session on Demographic Projections, Perugia, 3-7 May 1999.

de Groot, A.D. 1965. *Thought and Choice in Chess*. The Hague: Mouton.

Denzin, N.K. and Y.S. Lincoln, Eds. 1998. *The Landscape of Qualitative Research*. Thousand Oaks: Sage.

Elo, A.E. 1978. *The Ratings of Chess Players: Past and Present*. London: Batsford.

Ericsson, K.A. and P. Delaney. 1998. Working memory and expert performance. Pages 93-114 in R. Logie and K. Gilhooly (eds.), *Working Memory and Thinking. Current Issues in Thinking and Reasoning*. Hove, UK: Psychology Press.

Ericsson, K.A. and P. Delaney. 1999. Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. Pages 257-297 in M. Akira and P. Shah (eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York: Cambridge University Press.

Ericsson, K.A. and W. Kintsch. 1995. Long-term working memory. *Psychological Review* 102:211-245.

Ericsson, K.A. and A.C. Lehman. 1996. Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology* 47:273-305.

Hahn, E.J., C.P. Tourney, M.K. Rayens, and C.A. McCoy. 1999. Kentucky legislators' view on tobacco policy. *American Journal of Preventive Medicine* 16:81-88.

Hamilton, D. 1994. Traditions, preferences, and postures in applied qualitative research. Pages 60-69 in N.K. Denzin and Y.S. Lincoln (eds.), *Handbook of Qualitative Research*. London: Sage.

Hamilton, D. 1998. Traditions, preferences, and postures in applied qualitative research. Pages 111-129 in N.K. Denzin and Y.S. Lincoln (eds.), *The Landscape of Qualitative Research*. Thousand Oaks: Sage.

Hayes, J. 1981. *The Complete Problem Solver*. Philadelphia: The Franklin Institute Press.

Johnson-Laird, P. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Katona, O. 1940. *Organising and Memorizing*. New York: Columbia University Press.

Keilman, N. and A. Hetland. 1999. Simulated confidence intervals for future period and cohort fertility in Norway. Paper presented at the Joint ECE-Eurostat Work Session on Demographic Projections, Perugia, 3-7 May 1999.

Keyfitz, N. 1985. A probability representation of future population. *Zeitschrift für Bevölkerungswissenschaft* 11:179-191.

Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lee, R.D. 1993. Modeling and forecasting the time series of US fertility: Age patterns, range, and ultimate level. *International Journal of Forecasting* 9:187-202.

Lee, R.D. 1999. Probabilistic approaches to population forecasting. Pages 156-190 in W. Lutz, J.W. Vaupel, and D.A. Ahlburg (eds.), *Frontiers of Population Forecasting*. A supplement to *Population and Development Review*, Vol. 24, 1998. New York: The Population Council.

Lee, R.D. and L. Carter. 1992. Modeling and forecasting the time series of US mortality. *Journal of the American Statistical Association* 87(419):659-671.

Lee, R.D. and S. Tuljapurkar. 1994. Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association* 89(428):1175-1189.

Linstone, H.A. and M. Turoff. 1975. *The Delphi Method: Techniques and Applications*. Reading, MA: Addison-Wesley.

Lutz, W., Ed. 1996. *The Future Population of the World. What Can We Assume Today?* London: Earthscan, revised edition.

Lutz, W. and S. Scherbov. 1998. An expert-based framework for probabilistic national population projections: The example of Austria. *European Journal of Population* 14(1):1-17.

Lutz, W., W.C. Sanderson, and S. Scherbov. 1997. Doubling of world population unlikely. *Nature* 387(6635):803-805.

Lutz, W., W.C. Sanderson, and S. Scherbov. 1999. Expert-based probabilistic population projections. Pages 139-155 in W. Lutz, J.W. Vaupel, and D.A. Ahlburg (eds.), *Frontiers of Population Forecasting*. A supplement to *Population and Development Review*, Vol. 24, 1998. New York: The Population Council.

Lutz, W., J.W. Vaupel, and D.A. Ahlburg, Eds. 1999. *Frontiers of Population Forecasting*. A supplement to *Population and Development Review*, Vol. 24, 1998. New York: The Population Council.

Martino, J. 1983. *Technological Forecasting for Decision Making*. New York: North Holland.

Naylor, S.N. 1997. *Practical Reasoning in Natural Language*. Upper Saddle River, NJ: Prentice-Hall.

Parenté, F.J. and J.K. Anderson Parenté. 1987. Delphi inquiry systems. Pages 129-156 in G. Wright and P. Ayton (eds.), *Judgmental Forecasting*. New York: Wiley.

Pflaumer, P. 1988. Confidence intervals for population projections based on Monte Carlo methods. *International Journal of Forecasting* 4:135-142.

Popper, K.R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Reason, J. 1990. *Human Error*. Cambridge: Cambridge University Press.

Saariluoma, P. 1985. Chess players' intake of task relevant cues. *Memory and Cognition* 13:385-391.

Saariluoma, P. 1992. Error in chess: Apperception restructuring view. *Psychological Research* 54:17-26.

Saariluoma, P. 1995. *Chess Players' Thinking*. London: Routledge.

Saariluoma, P. 1997. *Foundational Analysis*. London: Routledge.

Sanderson, W.C. 1995. Predictability, complexity, and catastrophe in a collapsible model of population, development, and environment interactions. *Mathematical Population Studies* 5(3):259-279.

Stoto, M. 1983. The accuracy of population projections. *Journal of the American Statistical Association* 78(381):13-20.

Tuljapurkar, S. 1997. Taking the measure of uncertainty. *Nature* 387:760-761.

Vicente, K.J. 1988. Adapting the memory recall paradigm to evaluate interfaces. *Acta Psychologica* 69:249-278.

Vicente, K.J. and J.H. Wang 1998. An ecological theory of expertise effects in memory recall. *Psychological Review* 105:33-57.