WORKING PAPER
MANSHOLT GRADUATE SCHOOL

Combining expert opinion

Eligius Hendrix, Marcel Kornelis, Suzanne Pegge, and Michiel van Galen

DISCUSSION PAPER No. 25
2006

Mansholt Graduate School

Hollandseweg 1, 6706 KN Wageningen,
The Netherlands
Phone: +31 317 48 41 26
Fax: +31 317 48 47 63
Internet: http://www.mansholt.wur.nl/
e-mail: office.mansholt@wur.nl

Working Papers are interim reports on work of Mansholt Graduate School (MGS) and have received only limited reviews[1]. Each paper is refereed by one member of the Editorial Board and one member outside the board. Views or opinions expressed in them do not necessarily represent those of the Mansholt Graduate School.

The Mansholt Graduate School's researchers are based in three departments: 'Social Sciences', 'Environmental Sciences' and 'Agrotechnology and Food sciences' and two institutes: 'LEI, Agricultural Economics Research Institute' and 'Alterra, Research Institute for the Green World'. In total Mansholt Graduate School comprises about 250 researchers.

Mansholt Graduate School is specialised in social scientific analyses of the rural areas and the agri- and food chains. The Graduate School is known for its disciplinary and interdisciplinary work on theoretical and empirical issues concerning the transformation of agriculture, rural areas and chains towards multifunctionality and sustainability.

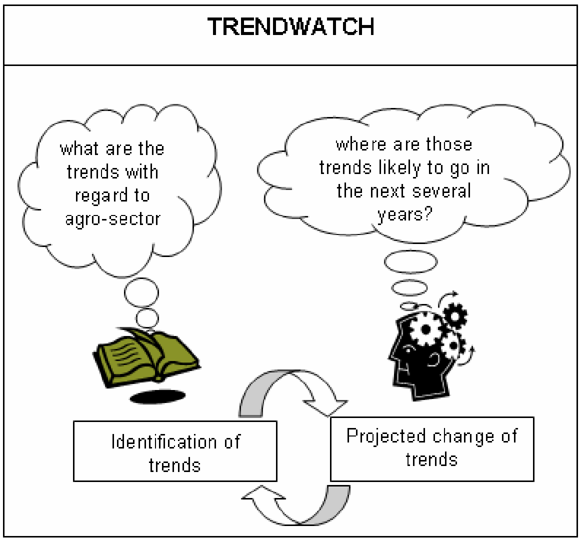
Comments on the Working Papers are welcome and should be addressed directly to the author(s).

Eligius Hendrix

Operations Research and Logistics Group, Wageningen University and Research Centre, Hollandseweg 1, 6706 KN Wageningen, The Netherlands.

[1] Working papers may have been submitted to other journals and have entered a journal's review process. Should the journal decide to publish the article the paper no longer will have the status of a Mansholt Working Paper and will be withdrawn from the Mansholt Graduate School's website. From then on a link will be made to the journal in question referring to the published work and its proper citation.

Combining expert opinion



Eligius Hendrix[1], Marcel Kornelis, Suzanne Pegge, and Michiel van Galen

Social Sciences Group, Wageningen University and Research Centers, The Netherlands

23 Januari 2006

This report is a result of the project TRENDWATCH, which has been financed by KB/LNV, ministry of Agriculture and Fisheries, The Netherlands

[1] Address correspondence to Eligius Hendrix, Operations Research and Logistics, Social Sciences Group, Wageningen University and Research Centre, Hollandseweg 1, 6706 KN, Wageningen

Abstract

In this study, focus is on a systematic way to detect future changes in trends that may affect the dynamics in the agro-food sector, and on the combination of opinions of experts. For the combination of expert opinions, the usefulness of multilevel models is investigated. Bayesian data analysis is used to obtain parameter estimates. The approach is illustrated by two case studies. The results are promising, but the procedures are just a first step into an appropriate combination of expert combination, which has to be completed on important issues, such as the identification of some well-known biases.

Key words: expert opinion, combining expert opinion, Bayesian data analysis, multilevel models, maximum entropy.

1. Introduction

The agro-food sector is embedded in a dynamic market environment. The world market dynamics on one side, and the behaviour of consumers on the other have a dramatic impact on the potential profitability of the sector. It is well-known that observed patterns in the market environment are not necessarily replicated into the future. Many events, such as taste changes of consumers, health awareness, obesitas etc., may change the long-run trend growth of agro-food markets. In order to anticipate on the ultimate effect of such market interventions or gradually changing slopes, it is useful to have an indication of the relevance of potential driving forces in the market. Therefore, decision makers such as politicians, chain directors, entrepreneurs, farmers, auction managers require information on speed and impact of possible changing trends in order to react pro-actively.

The *identification* of changes in the long-run growth path is a crucial first step in such a process. In reality, factors that affect market performance never occur in isolation. Instead, the long-run trends are affected by a combination of factors, which makes the identecation process extremely difficult. Detecting trends has several aspects. First, one should identify which factors may be labelled as an important potential change in the market performance. Second, one should find out their exact impact on the sector, and try to quantify how fast the changes are really developing over time. Third, decision makers should have an indication of the relevence of the changing developments over time, because, if the impact of market changing phenomena dies out over time, decision makers should not overreact.

Quantitative information is not always available from observed data and then one depends entirely on the opinion of experts. Experts and/or research institutes may differ in their assessment of factors that influence the agro-food sector. The *combination* of expert opinion may benefit the decision making process of market decisision makers. In addition, information on the possible *(dis)agreement* among experts may also help decision makers, as it indicates the level of uncertainty that surrounds projections of future events.

The development of a user friendly agro-outlook of combined expert opinions concerning environmental issues that potentially affect agro markets would be a useful tool for policy and market decision makers. The information obtained from individual experts should be combined in such a way, that the results are quickly obtained and easy to communicate. In addition, the underlying systematics, which give the ultimate outlook, should be adaptive to new market circumstances, as the markets' environment has become more and more dynamic.

In this study, we focus on the detection of future changes in trends that may affect the dynamics in the agro-food sector, and on the combinaton of opinions of experts in a systematic way. More specifically, we answer the following research questions:

- Which methods exist to elicit influencing factors from group sessions?
- What methods and concepts exist to combine expert opinion?
- How to implement methods in such a way that they are accepted and used by stakeholders?

The research questions are approached in several ways. For the first question, we briefly discuss several methods that can be used in group sessions, and we consider a real life example that has been developed by Agrotechnology & Food Innovations B.V. In addition, we give some limitation and drawbacks. To answer the second question, an extensive literature study has been performed to get an overview of existing methods of combining expert opinion. The viewpoint is to detect a method that can easily interpret and does not require too many data on the quality of experts. From the overview, a method is developed and elaborated. For the last question, we provide two examples, and discuss the success factors of model implications that have been formulated by Little (1970).

In Chapter 2, a procedure for the identification of trends is sketched. Chapter 3 describes the literature review on combining expert opinions. In Chapter 4, developed methods are described. In Chapter 5, the methods are applied to test cases to illustrate their ability and to detect characteristics of the methods. In Chapter 6 success criteria for systematic implemantation are discussed. Conclusions are given in Chapter 7.

2. Identification of expert opinion on trend changes

Making decisions under future uncertainty is unavoidable in everyday life. For instance, farmers have to decide whether it is better to harvest the next day or the next week, managers have to decide whether or not to introduce a new product, and students have to decide whether or not to concentrate on specific topics for an exam. It is impossible to perfectly foresee all these future events, and decision makers often try to reduce the degree of future uncertainty by using additional information. This anticipating behaviour is part of human nature. Beside models and heuristics, consulting *experts* is a possible instrument to reduce future uncertainty (for an extensive overview of the role of mathematical models and expert opinion in decision making processes, see Kornelis and Van Meijl, 2004).

In this chapter, we first discuss theoretical issues concerning the identification of expert opinion, followed by an example of such identification in practice. Finally, we consider some limitations and possible improvements on the identification of expert opinion.

2.1 Identifying expert opinion

The opinion of experts may provide important information for the decision making process of policy makers, managers, and/or may serve as input to a model. Particularly if the availability of "hard data" is limited, the opinions of experts can provide useful information regarding relevant uncertainties. The applications area of expert opinion is broad and includes nuclear engineering, aerospace, economics, meteorology, military intelligence, seismic risk, and environmental risk from toxic chemicals (for an extensive overview, see Cooke 1991).

Experts can not only interpret data characteristics, but also the underlying data-generating process, which cannot directly be observed. In general, the information is subjective and qualitative, and the investigators have to determine how to incorporate it into the projection-generating mechanism (Newbold and Bos 1994, pp. 7-8).

Two well-known systematically approaches of expert judgment are (i) the Delphi method, and (ii) cross-impact analysis. The Delphi method, developed at the Rand Corporation, is discussed in detail in Helmer (1968), Newbold and Bos (1994, p. 486), and Armstrong (1978 p. 108), among others. The Delphi-method is an iterative process of several rounds, in which a panel of experts answers 'if' or 'when' questions. These experts do not meet, but receive feedback from previous rounds. Consensus is not essential, because, it is also valuable to know on what aspects there is disagreement. It is difficult to evaluate the value of the Delphi method, because (i) most applications look many years ahead, (ii) the costs of performing the method for the short run is very high, and (iii) it is difficult to compare it with alternative approaches (Newbold and Bos 1994, p. 487). All in all, we consider the Delphi method as too complex to evaluate and too costly to perform.

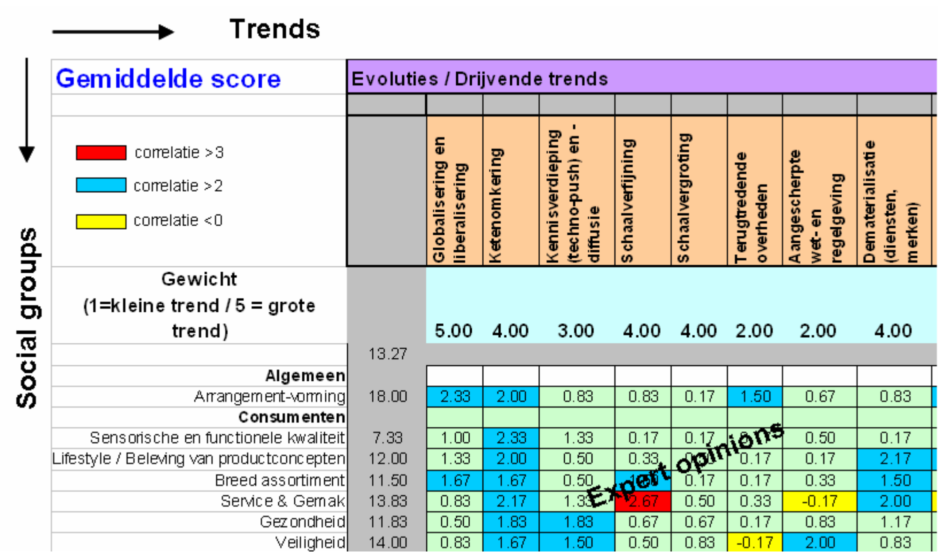
The cross-impact approach is often used in association with Delphi (see Linstone and Turoff 1975, and Helmer 1977). It assesses the interdependence of uncertain future events. An example of a cross-impact analysis question may be: "If A happens, what is the probability that B happens?". It is very difficult to anticipate future *causal* relationships (Newbold and Bos 1994, p. 488), therefore, we do not propose this type of questions.

Other methodologies on expert judgment include (i) visual inspection, (ii) personal interview, (iii) telephone interview, (iv) mail questionnaire, (v) informal meeting, (vi) formal meeting, (vii) group depth interview, and (viii) role playing (see Armstrong 1978, p.122 among others).

2.2 Identifying expert opinion in practice

An illustrative example of the identification of expert opinion in practice is the so-called "*Kenniskaart"* tool that has been developed at Agrotechnology & Food Innovations B.V. in Wageningen. This systematic approach makes use of a selection of publications, including books, magazines and websites that collect and publish breaks in trends. In addition, it makes use of experts who associate the effects of general trends with phenomena that are specific for the agro-sector. By offering insight in how trends affect consumer behaviour, how they evolved in recent years and how they are likely to evolve in the future, the "*Kenniskaart"* creates value for itself and other stakeholders What is not taken into account in this study, but would be of value for a *Trendwatch* is a distinction between various levels of aggregation. For example, "improving health" is a gradually changing trend on a global level, while "an increased demand for organic food and drinks in Dutch supermarkets" is a gradually changing trend at the meso/sector level.

The outcome of sessions with experts, that judge the selected agro-sector phenomena and driving forces, is a so-called *score card*. Figure 2.1. gives an example of such a score-card, where expert opinions are used to estimate the impact of trend breaks on future research in the field of agro logistics. In this study, experts were asked to give a score from 1 (no impact) to 5 (large impact) to the trends and the effect on a specific social group. Notice, that this is a multinomial scale. The ways in which expert opinion is combined depends crucially on the underlying scale on which the opinions are measured. We return to this issue in the next section. The average score of the experts are shown in the matrix in Figure 2.1. The colour indicates the correlation of the score among the experts.



Gemiddelde score		Globalisering en liberalisering	Ketenomkering	Kennisverdieping (techno-push) en -diffusie	Schaalverfijning	Schaalvergroting	Terugtredende overheden	Aangescherpte wet- en regelgeving	Dematerialisatie (diensten, merken)
Gewicht (1=kleine trend / 5 = grote trend)		5.00	4.00	3.00	4.00	4.00	2.00	2.00	4.00
	13.27								
Algemeen									
Arrangement-vorming	18.00	2.33	2.00	0.83	0.83	0.17	1.50	0.67	0.83
Consumenten									
Sensorische en functionele kwaliteit	7.33	1.00	2.33	1.33	0.17	0.17	[illegible]	0.50	0.17
Lifestyle / Beleving van productconcepten	12.00	1.33	2.00	0.50	0.33	[illegible]	0.17	0.17	2.17
Breed assortiment	11.50	1.67	1.67	0.50	[illegible]	0.17	0.17	0.33	1.50
Service & Gemak	13.83	0.83	2.17	1.33	2.67	0.50	0.33	-0.17	2.00
Gezondheid	11.83	0.50	1.83	1.83	0.67	0.67	0.17	0.83	1.17
Veiligheid	14.00	0.83	1.67	1.50	0.50	0.83	-0.17	2.00	0.83

Figure 2.1 an impact score card based on expert opinion

2.3 Limitations, problems and biases concerning expert opinion

A simple advantage of the use of expert opinions for the identification of trends can be observed from the exercise in Section 2.2. In order to measure data empirically, their definition, dimension and availability has to be assessed in an exact way. By the use of expert opinion, one can obtain information a priori without structuring into observable indicators. In this way, one gets an early identification of possible effects in cases that data are not available. Two major problems of expert judgment are the high costs (salaries), and the many possible biases. Below, we list some of the most important biases:

- Groupthink. Experts become supportive of each other and their 'leader', and therefore avoid conflicting opinions during their meetings (Janis 1972);
- Optimism. Experts not only reflect what they *think* that will happen, but also what they *hope* that will happen. For an example, see Hayes (1936);
- Anchoring. The way in which the question is asked anchors the response. For an example, see Tversky (1974)[2];
- Overreaction. People tend to overreact recent unexpected and dramatic events (De Bondt and Thaler 1985);
- Personal involvement. Increasing personal involvement with a future event tends to invoke a greater feeling of certainty that the event will or will not occur (Wright and Ayton 1989);
- Advocacy. Experts involved in the future outcome of the survey may overstate its virtues (see Tyebjee 1987, among others);
- Perceived controllability. If an expert feels that the occurrence of events is under his or her control, this perceived controllability results in increased overconfidence of the judgment (Wright and Ayton 1989).

Whatever the choice is of a certain methodology, the investigators have to deal with these types of biases.

Another issue is the combination of expert opinion. In the above example, the combined judgment was based on a simple arithmetic mean. Although such a procedure is fast and probably a good first step in the expert combining process, it does cannot deal with the aforementioned biases that potentially influence the outcomes. Indeed, to finally obtain a single outlook, the opinions of the experts have to be combined. The next chapter describes the investigation on methods for combining expert opinion. Concepts from literature have been investigated and methods have been derived that are of user friendly and can be interpreted easily.

[2] Tversky (1974) asked two distinct groups to quickly give an estimate of the outcome of $1\times2\times3\times4\times5\times6\times7\times8$, and $8\times7\times6\times5\times4\times3\times2\times1$, respectively. Although, the answer is the same (this is: 40320), the median estimate of the first group was 512, whereas the median estimate of the second group was 2250. Hence, the structure of the given question (descending or ascending, respectively) *anchored* the response.

3. Combining expert opinion, a literature review

In many decision making and model building processes, *multiple* experts are asked to give their opinions. The motivation for the use of more than one expert is the wish to obtain as much information as possible (Clemen and Winkler 1999). For decision makers and model assessors it is necessary to combine the opinion of various experts in such a way that it is beneficial to the process of interest. An important issue in this combination of expert opinion is the resolution of conflicting information or opinions. If there is disagreement among some experts, it is not trivial how one should deal with differences in competence and/or inter-dependencies among the experts.

Clemen and Winkler (1999) classify two streams of combination procedures, viz. mathematical and behavioral approaches. In this present study, we focus on mathematical procedures. Mathematical procedures can be further classified into axiomatic and Bayesian approaches.

Axiomatic approach. An axiomatic approach requires that the inter-dependencies between individual expert opinions and the combination of expert opinion satisfies a given set of axioms. The chosen set of axioms implies a particular combination rule. These rules typically lead to some form of a weighted mean. Examples are linear and multiplicative combinations of expert opinion.

Bayesian approach. In Bayesian data analysis, prior beliefs in the form of probability statements are multiplied with likelihood functions[3] to obtain a joint probability distribution. From this joint probability distribution, posterior distributions for the parameters of interest can be derived. In the setting of our focal interest, this Bayes rule applies as follows: the decision maker updates his or her a priori beliefs on an event of focal interest, using a likelihood function that is associated to the opinion of experts (for a similar reasoning, see French 1985, Lindley 1985, and Genest and Zidek 1986). The Bayesian data analysis can be very complex in its application. The practical use has increased in the last decade, as more and more Bayesian analyses make use of available simulation methods, such as Markov Chain Monte Carlo simulation, available in software packages. An example of a software package that estimates distribution function parameters using these simulation techniques is WinBugs (WinBugs 2003). Note that Bayesian data analysis is not restricted to the combination of expert options, but it is a fundamental statistical framework. For detailed introductions into this field of statistics, we refer to Berger (1985), Gelman et al. (1998) and Lee (1997).

[3] Formally, a likelihood function is defined as follows: 'For observed data, x, the function $L(\theta) = f(x|\theta)$, considered as a function of θ, is called the likelihood function' (Berger, 1985, p. 27).

3.1 An example of the axiomatic approach and its limitations.

In their study, Myung et al. (1996) use the so-called maximum-entropy approach as a tool to combine the opinion of experts. The intuition behind the maximum entropy approach is as follows. Suppose that the decision maker faces N possible situations, and the experts are asked to choose, out of these N situations, the one situation which is the most likely to occur (in their opinion). The opinion of the experts take the form of a probability distribution ('In my opinion, the probability that situation j occurs is equal to p_j'). The entropy measure of a probability distribution is maximum if all possible situations are equally likely to occur. In our example, the associated probability for each situation equally likely to occur is $p_j = N^{-1}$, which is the geometric mean of the N situations. This outcome corresponds to a situation of *maximum uncertainty*. In other words, if an expert is most uncertain about which situation is the most likely to occur, he or she is expected to report the same probability p_j for each of the N situations. Consequently, any relevant information about the probable occurrence of one of the situations, would affect this uniform probability distribution, and thus reduce the maximum uncertainty. This reduction of uncertainty is represented by a drop in the value of the entropy measure. The uncertainty is zero if one of the possible situations is actually observed. In that case, its associated probability of that occurred situation is equal to one, while all the other situations have a probability of zero. Indeed, the entropy becomes zero as well. Therefore, we can interpret entropy as a measure for average uncertainty. It can be shown that if information about the probability distribution is available in the form of moments, the solution that maximizes the entropy given these moments, utilizes all information contained in the set of moments, but is mostly non-committal with respect to the unknown information (Jaynes 1957). Note that these moments act as restrictions in an optimization problem. Myung et al. (1996) show how (i) competence, (ii) inter-dependencies, and (iii) past behavior in similar situations, can enter the optimization framework as moment restrictions. In an artificial example, they investigate how changes in the available information influence the entropy measure and thus the combination of the expert opinions.

Two important aspects of the approach in Myung et al. (1996) are worth mentioning. The first one is the question where to obtain the information on (i) competence, (ii) inter-dependencies, and (iii) behavior in similar situations. Myung et al. (1996) suggest using information from *past* data. Often, however, past data is not available. In such situations, the decision-maker or investigator can subjectively generate a priori distributions for the parameters and moment constraints. Some efforts have been directed toward the development of Bayesian combination methods that are suitable for the use of subjective judgment in determining the likelihood function. An example in this field is given in Lipscomb et al. (1998). The second issue is a major limitation of the maximum entropy approach as presented in Myung et al. (1996). If the number of experts increases, the number of potential combinations, that have to be considered, grows exponentially. To illustrate this, say that fifty experts are asked to judge the probable occurrence of twenty situations. The corresponding probability has to be determined at 2.25E+66 possible outcomes. It is problematic for numerical algorithms to solve such large problems in polynomial time. We see two ways which may be helpful in this context.

First, one may consider the use of simulation methods, as have been developed in the area of Bayesian statistics, such as Markov Chain Monte Carlo methods, to generate outcomes for the maximum entropy approach. We see this as an interesting topic for future research. Second, we argue that the available information *about* the experts may enter the combination problem at a higher level of hierarchy than the information given *by* the experts. Therefore, the use of multiple levels may also help to reduce the computational burden of the above discussed entropy approach.

3.2 Multilevel models.

In many areas, including economical, social, and biological sciences, multilevel or hierarchically structured data can be found. Examples of hierarchically structured observations are the opinions of experts, in which the individual experts are subject to the influences of grouping. For example, experts work in business units or research groups; research groups are within research centers, and research centers may be within universities.

In such a dataset, one can identify four different levels of hierarchy, and a multilevel approach would assign the experts to level one, research groups to level two, research centers to level three, and universities to level four. Units at one level are considered as being grouped, or nested within units at the next higher level. The parameters of interest, e.g. a probability statement by an expert, can be seen as related in some way by this hierarchical structure. This implies that the inter-dependence among them should be reflected in a joint probability model. In other words, the opinion of experts at research group j should be related to each other. Such hierarchical thinking may help to understand and solve multiple-parameter problems and, in our setting, may be beneficial in the development of computational strategies for the combination of expert opinions.

In a Bayesian setting, the inter-dependency is achieved in a natural way by using a prior distribution in which, for example, the opinions of experts at research group j are viewed as a sample from a *common* population distribution. In their turn, the various common distributions are also viewed as inter-related, but at a higher level. In this way, the data set is modelled hierarchically, because the observable expert opinions are modelled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as hyperparameters. Note that the observed data can be used to estimate the hyperparameters, even tough the values of these hyperparameters are not observed themselves.

3.3 Scales.

It is important to note that the scales which are used to measure the opinion of the experts determine the appropriate combination technique to a large extent. In this study, we categorize between scales for continuous and categorical responses. Examples of continuous response data are the expert opinions on prices, waiting time, or future revenues. Examples of categorical response data include the choice for a certain product, policy, job candidate, or a dichotomized quality statement. If responses on a continuous scale are regarded as outcomes of a probability distribution, it is natural to consider a continuous density function, e.g. the normal density function, as the underlying data generating process. In the case of categorical data, the researcher may consider discrete density functions, e.g. the binomial (when there are two categories), the multinomial (when there are more than two categories), or the Poisson density function. As we will see in the methodology section, in a Bayesian data analysis, the choice for a certain density function is a crucial aspect of the combination process, because it may lead, for example, to different conjugate prior distributions.

4. Methodology

In this study, we consider the use of multilevel models for the combination of expert opinion. More specifically, we perform a Bayesian data analysis, where we account for the potential hierarchical structure in the observed opinion of experts. As aforementioned, the underlying date generating process of the opinion of experts may consist of multiple levels. The number of these levels represents a type of inter-dependency between experts. For example, if two experts belong to the same research group, they may show a stronger inter-dependency in their opinions as compared to experts from two different research groups. Measuring such inter-relationships can provide valuable information for the decision making process of focal interest, as it gives an indication about the (dis)agreement among experts. In this study, we consider two levels of hierarchy. We label the first level as 'individual level'. At this level, the opinion of each individual expert is observed. We label the second level as 'group level'. At this level, we assume that the opinions of individual experts are drawings from a common distribution function. So, we assume that all experts belong to the same group, and we do not consider multiple sub-groups of experts. Using the hyperparameters of the 'group level', we can obtain the desired combination measure of the opinions of experts. In this section, we first discuss opinion of experts given on a continuous scale, and then we discuss the combinations of observations on a categorical scale.

4.1 Expert opinion measured on a continuous scale

Lipscomb et al. (1998) (LPH) provide a multilevel approach for the opinion of experts that is measured on a continuous scale. In their empirical application, medical experts were asked to give their judgment about the mean amount of time that was required to perform a certain medical care service where 'time' is measured on a continuous scale. For the continuous case, we follow the procedure of LPH.

At the 'individual level', the opinion for each of the I experts $(i = 1,..,I)$ is modelled as a drawing from a normal distribution with mean μ_i and variance σ_i^2. This normal distribution reflects the subjective probability judgment of expert i. In LPH, the mean and the variance were asked from the experts themselves, but other options are possible as well. One may think about deriving the mean from a survey in which the same opinion is asked in different ways, in order to obtain a variance or let the investigator a priori decide on the value of the variance. So, at the individual expert level, we have:

$$y_i | \mu_i \sim \mathrm{N}(\mu_i, \sigma_i^2), \text{ where } i = 1,...,I \tag{1}$$

where y_i is the observed opinion of expert i . For example y_i may represent a judgement for the mean duration of a medical service in the empirical application of LPH.

The parameter of interest is μ_i as that one should be combined towards one value. Since each expert belongs to the same group that has to judge on a common value of the duration of the treatment, it is assumed that these individual μ_i 's are drawings from a common distribution function representing this group. So, the 'group level' is associated with the 'individual level' via the μ_i 's ($i = 1,...,I$), as is indicated in Figure 4.1. Figure 4.1 graphically depicts the hierarchical structure of the expert opinions. The various parameters are presented in ovals, if they are a drawing from a distribution, and in squares if they are a priori known scalars. In general, σ_i^2 may also be drawn from a higher level distribution. However, in LPH they are assumed to be known and given by the experts themselves. We follow LPH in this matter, as σ_i^2 is not a parameter of our focal interest.

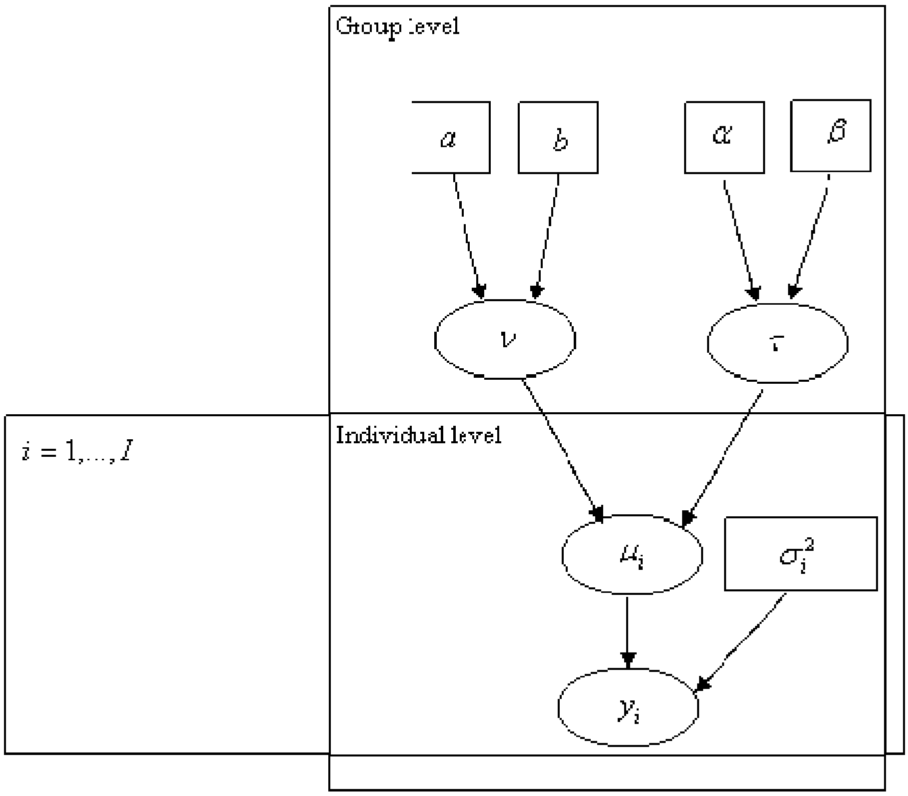


Figure 4.1Multilevel model about medical treatment duration as used by Lipscomb et al. 1998

The individual means of the normal distributions are supposed to be drawings from distributions at a higher level. These higher level distributions are characterized by unobserved hyperparameters. The common distribution function is again assumed to be normally distributed, now with mean ν and variance τ^2:

$$\mu_i | \nu, \tau \sim \mathrm{N}(\nu, \tau^2), \text{ where } i = 1,..., I \tag{2}$$

The "disappearance" of the index i in the normal distribution of (2) reflects the fact that this distribution belongs to the 'group' rather than the 'individual' level. The hyperparameter ν reflects the combined expert opinion of focal interest. LPH opt for a flat improper density on ν, which we elaborate as follows:

$$\nu \sim U(a, b), \tag{3}$$

where $U(,)$ is a uniform distribution, and a and b are a lower and upper bound. The parameters a and b are known and given by the researcher. The interpretation of (3) is that the common expert opinion is estimated and has a likely probability of occurring anywhere between a and b. The variability in the opinions among the experts is captured by the parameter τ^2. It is common to model this parameter by means of a gamma distribution for the inverse of τ, because that is a so-called conjugate distribution (Gelman et al. 1995, p. 480):

$$\frac{1}{\tau} \sim \Gamma(\alpha, \beta)\text{, for } \alpha > 0 \text{ and } \beta > 0 \tag{4}$$

where α and β are known parameters and given by the researcher. The parameter α is called the *shape*, and the parameter β is called the *inverse-scale* of the gamma distribution. Together they determine the expected value and variance of τ through:

$$\mathrm{E}[\tau] = \frac{\beta}{\alpha - 1}\text{, for } \alpha > 1, \tag{5}$$

and

$$\mathrm{var}[\tau] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\text{, for } \alpha > 2, \tag{6}$$

respectively. LPH identify the values for α and β as follows. The smallest integer of α that gives a finite variance is α=3. A low choice for alpha results in a heavy tailed distribution, so that extreme discrepancies between the prior and the data would be resolved in favor of the data. Therefore, LPH choose α=3. Expert opinion is then used to pinpoint the mean of τ (given in (5)), and finally the value for β is derived via (5).

The posterior distribution on ν represents the combination of the opinion of experts, where each opinion is weighted according to the accuracy, as captured by σ_i. The variance around ν can be interpreted as follows: the smaller the variance around ν, the larger the agreement among the experts. The Bayesian data analysis also provides posterior distributions for the individual means μ_i. These posterior distributions depend on y_i, but also on ν. Since ν is the group level mean, the dependence of y_i on ν reflects the underlying idea that the experts are inter-related. The variance of the group level mean represents the dependence among experts. If the posterior estimate of τ^2 is small relative to σ_i^2, the expert's posterior estimate of μ_i contains much shared information, while in the situation where τ^2 is large relative to σ_i^2, the location of the expert-specific posterior on μ_i will be close to the individual expert's own opinion y_i.

4.2 Expert opinion measured on a discrete scale

Suppose, that *I* experts are asked to give their opinion about the potential occurrence of J categories as follows. Each expert is allowed to divide one-hundred credits over the *J* categories in such a way that the scores of each category are positively related to the probable occurrence (probable in the view of the expert). So, if expert *i* believes that the item could belong to any category, he gives all *J* categories the same score and y_{ij} becomes $100 \cdot J^{-1}$ for each $j = 1,..., J$. If, in contrast, the expert believes that the item should really belong to only one category, expert *i* gives hundred credits to that one category, and zero to all others. (See for similar reasoning Hardaker et al. 1997, Chapter 3). In this way, we can obtain expert opinions for categorical data in the form of probability statements. A natural way to model such data is by means of the multinomial (*MN*) distribution. This implies that:

$$y_{ij} \mid p_j, n \sim MN(p_j, n), \tag{7}$$

Where n is the number of credits to be distributed over the categories and p_j, the group level probability an expert classifies the item to category *j*. In our setting, we are interested in the posterior mean of hyperparameter p_j. In the Bayesian literature, this hyperparameter often is assumed to follow the conjugate Dirichlet distribution:

$$p_j \mid \nu_j \sim Dirich(\nu_j), \tag{8}$$

where vector ν determines mean and variance of p_j. The difficulty of estimating the accuracy by which experts are able to identify category *j*, can be modelled by choosing parameter values of the uniform distribution ν_j is taken from :

$$\nu_j \sim U(a_j, b_j), \tag{9}$$

Such that the researcher can give values for a_j and b_j. Figure 4.2 graphically depicts the multinomial strategy.

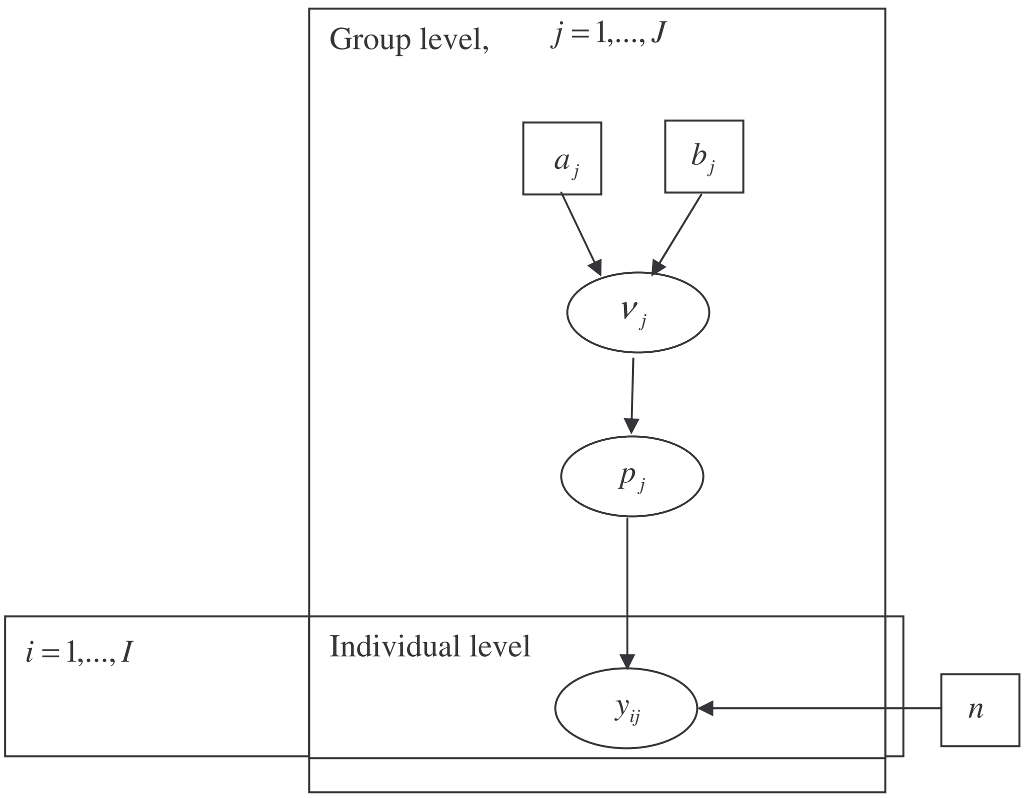


Figure 4.2 Multilevel model for data on categories

Drawback of this approach compared to the one in Section 4.1, is that the variance or accuracy of individual expert *i* is not modelled. An alternative approach may be a multilevel log-linear Poisson model. It is well known that the multinomial model is likelihood-equivalent to the Poisson log-linear model (see Palmgren 1981, McCullagh and Nelder 1989, Chap. 6, Lindsey 1995, Chap. 2, and Chen and Kuo 2001). The Poisson log-linear model can be extended to incorporate random effects, but can still be easily applied (Chen and Kuo

2001). Many existing software can fit the Poisson log-linear model, since it belongs to the generalized linear models family. At the individual expert level, we have:

$$y_{ij} \mid \lambda_{ij} \sim Po\left(\lambda_{ij}\right), \text{ where } i = 1,...,I, \quad j = 1,...,J, \tag{10}$$

where y_{ij} is the opinion of expert *i* that follows a Poisson distribution with mean and variance λ_{ij}, *I* is the total number of experts, and *J* is the total number of categories. At the 'group level', the parameter λ_{ij} is associated with the hyperparameters ν and τ as follows:

$$\log\left(\lambda_{\text{ij}}\right) \mid \nu_j, \tau_j \sim \text{N}\left(\nu_j, \tau_j^2\right), \text{ where } i = 1,...,I, \quad j = 1,...,J \tag{11}$$

Taking the logarithm of λ_{ij} is a common transformation in order to map the multichotomize parameter onto the continuous space. As a consequence, the normal distribution can be applied, and for the hyperparameters ν_j and τ_j everything has been defined before. The Poisson log-linear approach is graphically depicted in Figure 4.3.

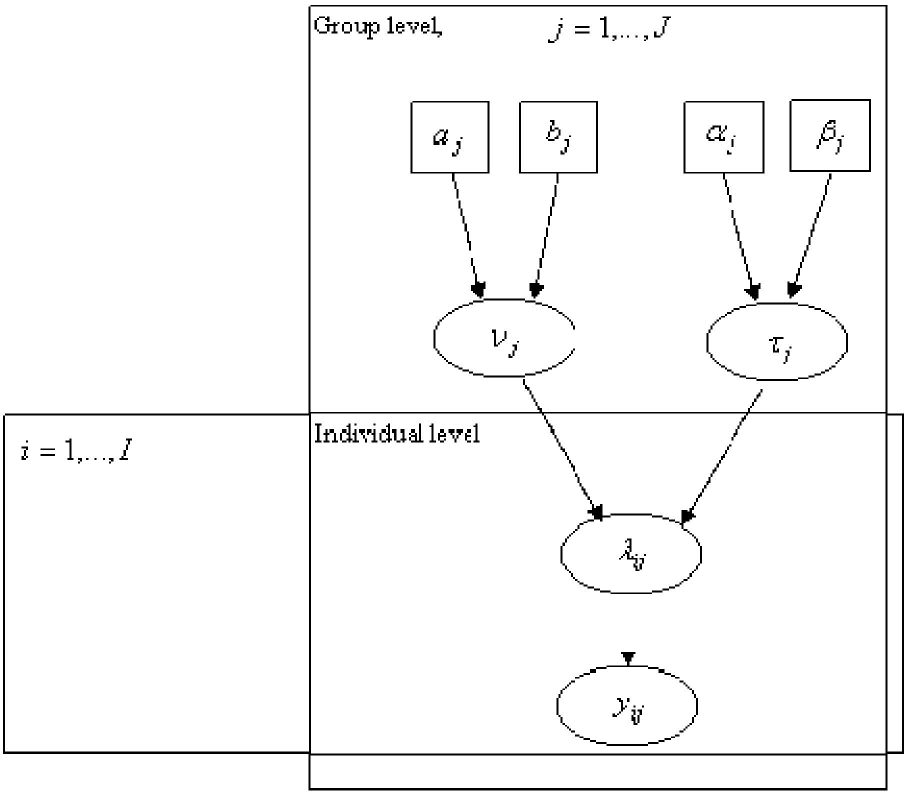


Figure 4.3 The Multilevel log-linear Poisson model

5. Two case studies

In this chapter, we discuss two case studies that make use of the obtained insights of the previous chapter. The first study obtains expert opinion on a discrete, and the second study on a continuous scale.

5.1 Estimating future yields of a tomato seedling[4]

Horticulture plays an important role in the Dutch economy, with a total production volume of 4.1 billion kilos and a production value of 7 thousand million (4% of gross domestic product) in 2002 (Van Klink and Visser, 2004). The horticultural sector consists of 55,000 firms (7% of the total number of firms in the Netherlands), which employ about 200,000 people and use up to 93,000 hectares of land, including 4,250 hectares with greenhouses (Greenery, 2004). Horticulture includes both the production of ornamentals and of edible crops.

Within the edible crop sector, tomatoes are the largest vegetable crop produced in the Netherlands and the second most favourite vegetable consumed (Productschap Tuinbouw, 2004), as is shown in Table 1. In 2004 the sector produced almost 650 million kilograms of tomatoes of which about 540 million kilograms of tomatoes were exported. The main markets for Dutch tomatoes are Germany and the United Kingdom (Ministerie LNV, 2004).

Table 5.1: Total agricultural activities in the Netherlands (2004),

	Share (%)	Production value € (thousand million)
Horticulture	41	6,9
Edible crops	***13***	***2,2***
Ornamental	*28*	*4,7*
Arable farming	13	2,2
Livestock production	25	4,2
Intensive livestock production	20	3,4

Source: Productschap Tuinbouw

In the tomato supply chain, breeding companies play an important role. The total tomato yield of each individual tomato plant depends, among other things, on the initial quality of the seedling. However, determining the quality of the seedling is a complex process. This is caused by the large natural variation in the plants and the broad range of sorting criteria. Seedlings have to be sorted on singular criteria concerning e.g. leaf area, stem length, and leaf curvature, but also on more complex issues concerning e.g. the likelihood that a plant is budless or rogue, and the regularity of the leaf shape (Koenderink et al, 2004). Because of its importance for the firm's potential productivity and, hence revenues, the pre-selection of seedlings is a highly important task, which needs highly skilled expertise.

Breeding companies use the opinion of experts to predict the future yields of a tomato seedling. Because it is a human task, subjective and personal factors will always play a role. In addition human expert-opinions are time-consuming and thus expensive. To overcome these problems research is done to create a computer-based expert system. Given its importance, it is not surprising, that there is a need to ask more than one expert for his or her

[4] The presented case forms a small part of the PhD research of Nicole J.J.P. Koenderink.

judgement. The relevant question then becomes, how to incorporate the opinion of multiple experts into a single expectation.

Tomato seedlings are sorted on the basis of their expected production potential when they are between 10 and 15 days old. The sorting specialists that assign the seedlings to their quality class work according to guidelines issued by the Netherlands Inspection Service for Horticulture. Figure 5.2 provides an illustrative example of the leaves that the experts have to judge. These guidelines require the young plants to be assigned to one of the following four categories:

- Plants that are of good quality;
- Plants that are of second choice quality;
- Plants that are too small;
- Plants that have defects, with indication of the occurring defect.

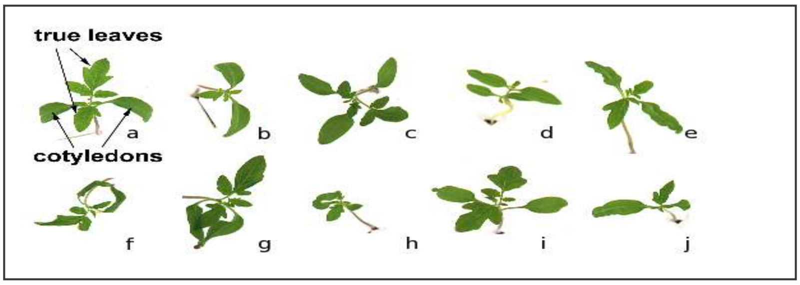


Figure 5.2: A set of seedlings that have to be assigned to the correct quality class.

Experts are asked to assign a probability judgement to the same uncertain event; the future yield of one tomato seedling A. In this case credits are used to represent relative probabilities. The possible alternative outcomes (good quality, second choice quality, too small, defects) are listed in a table. The experts are given 100 credits which they can allocate across the spaces in the table. When the expert is fully satisfied with his allocation of credits, probabilities can be calculated according to observed cell frequencies (Hardaker et al, 1997). Figure 5.3 shows the (hypothetical) answers of the 4 experts.

Seedling A	Good quality	Second choice quality	Too small	Defects
	Credits	*Credits*	*Credits*	*Credits*
Expert 1	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (25)	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (25)	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (25)	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (25)
Expert 2	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (75)	●●●●● ●●●●● (10)	●●●●● ●●●●● (10)	●●●●● (5)
Expert 3	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (60)	0	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (40)	0
Expert 4	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (30)	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (30)	●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● (30)	●●●●● ●●●●● (10)

Figure 5.3: Method for assessing the probabilities of future yields

In figures 5.4, the probability distribution for seedling A of each of the 4 experts is graphically represented. The peak of each individual expert indicates their perceived most likely outcome.

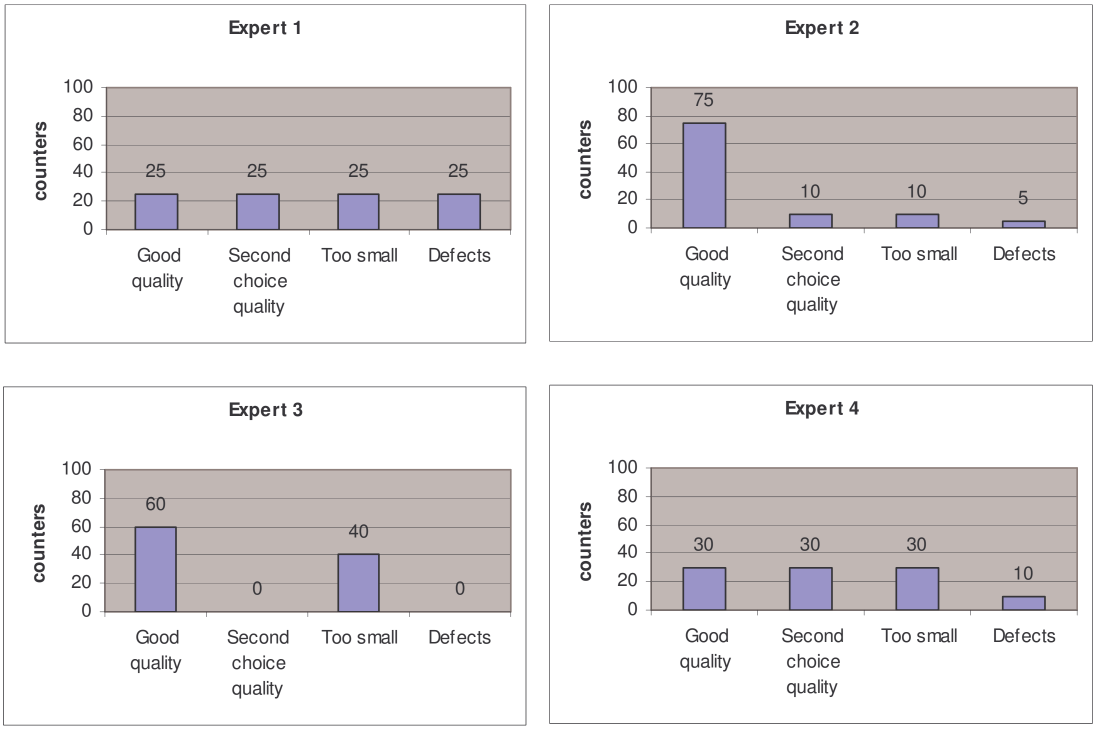


Figure 5.4: Expert opinion on future yield probability of seedling A

On the basis of these four individual probabilities it is difficult to determine the 'real' quality of the tomato seedling. To incorporate the opinion of multiple experts into a single expectation the multilevel model of Section 4.2 was estimated in WinBUGS (a software package). The outlook on the future tomato yield of seedling A is illustrated in Figure 5.5.

Figure 5.5: Estimated future yield probability of seedling A

The figure shows that it is most likely that seedling A has a good quality and hence will have a high tomato yield in the future. In this example the assumption is made that for an expert all four categories are equally difficult to determine.

If we now assume that within the four quality categories, it is more difficult to determine that a seedling has defects or has second choice quality than it is to determine that the seedling is of good quality or too small. If we apply this assumption to our model, the results are as illustrated in Figure 5.6.

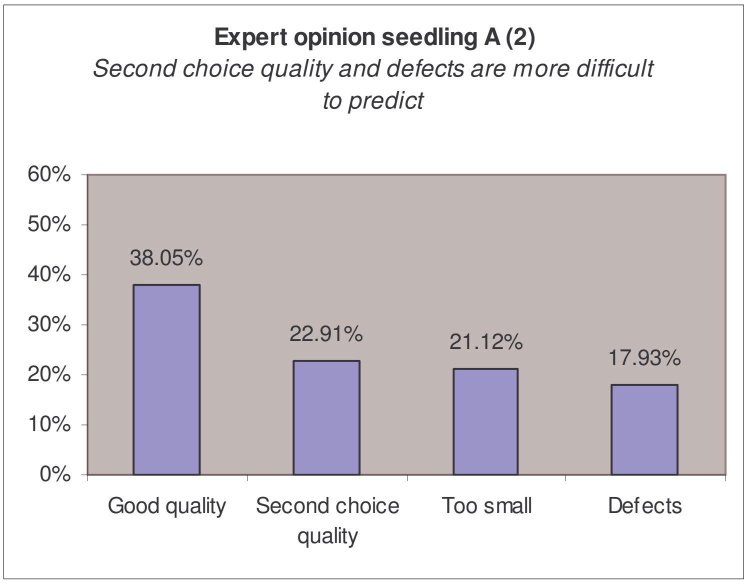


Figure 5.6: Estimated future yield probability of seedling A

The new figure shows that it is still most likely that seedling A has a good quality and hence will have a high tomato yield in the future. The chance that the seedling is too small has declined. The chance that the future outcome will be second choice quality or defects has increased.

To conclude, the examples above show how four single expert opinions about an uncertain future event, in this case the yield of a tomato seedling, can be aggregated to one opinion by using a multilevel model. Ideally we would like to add the reliability of each expert into the model. Is one expert always certain about his answers or does he distribute the credits more or less equally? Unfortunately the current used WinBUGS model is limited in the number of expert opinions and/or seedlings it can handle. To overcome this ‘problem’ might consider either buying a licence for a package than can handle more data or programming a special WUR-trendwatch model.

5.2 Aggregating expert opinion on changes in tomato prices

More than half a million people work in the Dutch agribusiness complex, of which 168 thousand are employed the primary sector. In terms of employment, horticulture is the largest primary subsector with 110 thousand employees (Tuinbouw). Its production value of almost eight billion euro represents almost 40 percent of primary sector production value in the Netherlands. The sector produces mainly for export markets, such as Germany and France. Horticultural products like tomatoes, sweet peppers, cauliflowers and tulips exhibit large dynamic price fluctuations, due to changing demand and supply conditions. Business decision-making, e.g. negotiating price and volume of transactions with potential customers, requires accurate price predictions.

Given its importance, it should come as no surprise that market decision makers develop strategies to reduce price uncertainty. Such strategies include among others, longer-term contracts to stabilize prices, price model forecasts, and expert judgements and consultation. This study concentrates on the application of expert judgements to predict future price levels. More specifically, we investigate the possibility of combining the opinions of multiple experts into a single prediction, or 'outlook', which is easier to communicate than a set of potentially conflicting opinions. The degree of agreement or disagreement among experts should be an important characteristic of a single outlook as it can be interpreted as a measure of 'between-expert' uncertainty about the true future price level (Lipscomb et al, 1998). Many factors affect this uncertainty including the actual variability of prices, the differences in expert quality of assessment and information, and e.g. strategic behaviour by panellists.

To combine expert judgement, hereby accounting for between-expert uncertainty in relation to the subjective uncertainty of each individual expert (denoted as 'within-expert' uncertainty), we make use of so-called random-effects models. In this setting, the hierarchical Bayesian approach is applied as outlined in Section 4.1, which is a very flexible way of dealing with within and between-expert uncertainty. This type of analysis has been used in e.g. assessing expert opinions in forecasting stock prices (Ying et al 2005) and in predicting record sales for new albums (Lee et al 2003). We compare the Bayesian approach with the arithmetic mean prediction and time-series models to illustrate its usefulness.

The model is applied on the Dutch tomato industry, as being one of the most important horticultural products. In particular, we are looking for a way of aggregating expert opinion about future price of Dutch bunched tomatoes, into one synthesis forecast. Bunched tomatoes are one of the major horticultural products in the Netherlands, whose prices can fluctuate substantially both within and between seasons. The ultimate experts on market prices for Dutch tomatoes are the people involved in selling and buying them on a day-to-day basis. In markets there are always contradicting interests. Sellers profit from high prices, buyers profit from low prices. Furthermore, tomato traders may be influenced by a common 'market sentiment' or by reading the same newspapers. Hence, the opinions of buyers and sellers may differ systematically.

The experiment in this case study is targeted towards finding a relatively easy way to pool buyers' and sellers' and other experts' opinions about future market developments. We investigate the following research questions by means of our tomato price case study:

Can we improve on the arithmetic mean prediction by adding information on the expert certitude and quality?

For the case-study experiment we ask three buyers of tomatoes in a particular Dutch auction and three sellers of tomatoes in the same market to fill out a short survey. The survey consists of two parts: a) a prediction of the price of bunched tomatoes, from last week until four weeks into the future, b) an indication of the experts certainty about his own predictions and an indication of the experts qualifications. *Question 1: Give your best estimate of the price of Dutch bunched tomatoes for the next 4 weeks, average weekly price, Class 1, in euro per kilogram. Question 2: How certain are you about your estimates for week 4? Could you give an upper and a lower price limit, such that you feel 95 percent confident that the true price will lie within the interval?*

We assume for the moment that the expert opinions are normally distributed with mean *m* and standard deviation s^2. Table 1 summarizes the answers (hypothetical) of the six experts, where the standard deviation is calculated from the upper and lower limit, under the assumption that the upper limit lies two standard deviations from the mean prediction. These distributions constitute the first level in the hierarchical model: the data. Each of the experts is considered to be independent of the other experts in making a judgment. Note that the arithmetic mean of the expert opinions is 1.48 euro per kilogram.

Table 5.2. Unpooled expert estimates of the price of bunched tomatoes in week 4

	Prediction	Lower limit	Upper limit	
Expert A	1.40	1.20	1.60	(0.10)
Expert B	1.60	1.30	1.90	(0.15)
Expert C	1.55	0.75	2.35	(0.40)
Expert D	0.75	0.65	0.85	(0.05)
Expert E	1.40	1.35	1.45	(0.05)
Expert F	2.20	0.20	4.20	(1.00)
Arithmetic mean	1.48			

Where figures in parenthesis are elicited standard deviations

Bayesian analysis estimates the posterior or synthesis probability distribution of tomato prices from the data and from prior information that we might have to our disposal. We might understand the opinion of each expert as a sample from the distribution of the true state of tomato price developments. The likelihood that the sample reflects the true state accurately depends on the distribution of the prior. “Each posterior mean reflects the extent of common information and agreement among experts” (Lipscomb et al, 1998)

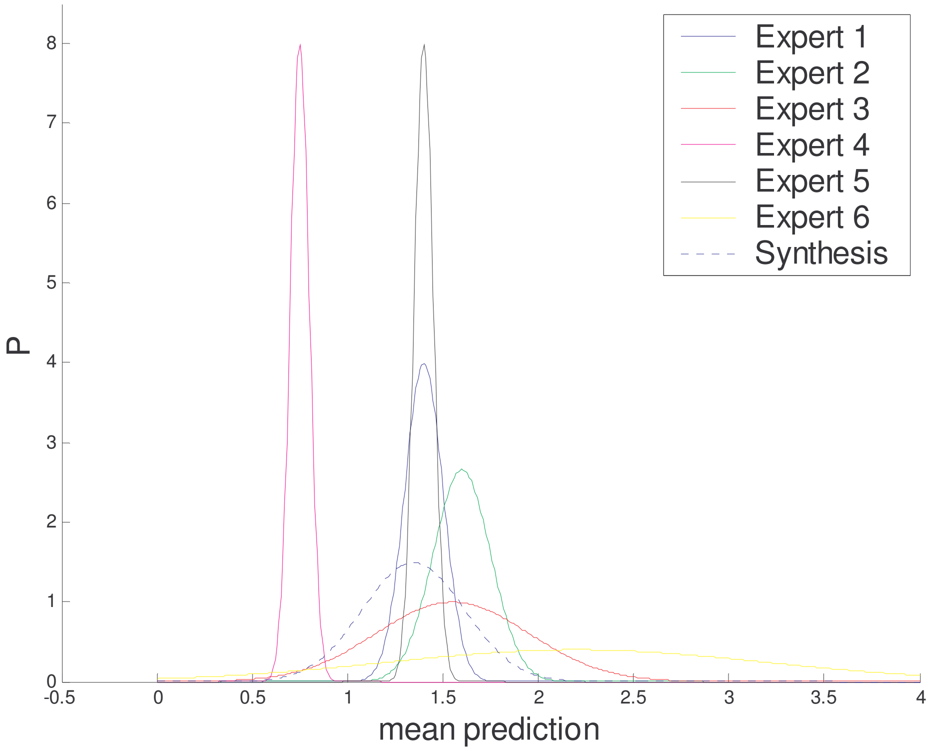


Figure 5.7 Probability distributions of expert opinions

Figure 5.7 shows induced normal probability distributions for each of the six experts. Under the influence of the prior values of ν and τ the synthesis distribution conveys more common information (prior) or more information from our observations. Where ν represents the mean of the synthesis distribution and τ the standard deviation. Prior ν is uniformly distributed between –100 and 100, indicating that we have little or no prior feeling about the true price. For the precision of μ, we use an informative prior τ, with an inverse gamma distribution in which the "degrees of freedom" or "shape" parameter α is set to 3 and the "scale" parameter β is 1. Higher values of β result in lower mean and standard deviation of τ. Consequently, the estimate for the posterior ν will be pulled more to the prior and contain less information from the experts. "The posterior distribution on ν combines information from all experts, naturally weighting each according to his/her accuracy, as captured by s_i. […] The greater the agreement among the experts, the greater the concentration around a central tendency, or "synthesis" estimate." (Lipscomb et al 1998)

Using the WinBugs modelling software, we estimated likelihood and posterior distributions for the experts' own opinions and the overall true state of the tomato price. Table 5.3 below presents the main results of our experiment. We can see from Table 5.3, that the posterior mean is 1.347; considerably lower than the arithmetic mean of 1.48. This result is driven by our choice of beta from the between-expert variation. Lower values of prior mean ν would have resulted in more acceptance of disagreement among experts, and consequently more emphasis on the observed data.

From Lipscomb et al (1998): "In general, the spread of the population distribution governs the dependence among experts. If the population variance τ^2 is small compared to the experts own variance – that is, if the posterior estimate of τ^2 is small relative to s_i^2 [as is true in our case] – the expert's posterior estimate of μ_i will contain much shared information." This effect is clear in the posterior mean for Expert F. The model judges that this expert must be mistaken and the likelihood of being right in his/her prediction is very small. Therefore, expert F's prediction is being 'shrunk' towards the shared mean. Expert D on the other hand is also far from the mean prediction, but remember that our prior $\nu \sim U(-100,100)$ says almost nothing about the mean of the true price. The mere fact that expert D is much more certain about his/her prediction ensures that it is not being 'pulled' towards the arithmetic mean.

Table 5.3. Pooled results

	Mean		Lower bound	Upper bound
Expert A	1.398	(0.099)	1.204	1.590
Expert B	1.579	(0.145)	1.294	1.865
Expert C	1.405	(0.484)	0.461	2.375
Expert D	0.756	(0.050)	0.658	0.853
Expert E	1.399	(0.050)	1.302	1.496
Expert F	1.543	(0.526)	0.543	2.627
Posterior ν	1.347	(0.266)	0.832	1.889

Values in parentheses are posterior standard deviations.
Based on an informative prior τ, with $\alpha = 3$ and $\beta = 1$.

We conclude that the use of Bayesian models and expert predictions is a very helpful and easy way to make funded assessments of future prices. Furthermore, this type of analysis could be applied to a wide range of problems. Especially complex stochastic processes for which little or no information is accessible can be analysed this way. Examples of such questions are: by how much will productivity increase over the next years?, will intensification in agricultural production increase or decrease?, will demand for a certain product increase or decrease?

6. Implementation of models[5]

Systematic approaches that support the decision making process are only effective if they match with the thinking and reasoning processes of the decision makers (see Van Bruggen and Wierenga 2000 for similar reasoning). Therefore, the system, and thus the models, will have to be appropriately implemented in the client's decision making process. Little (1970) provides six criteria for successful model implementation in the decision-making process (see Leeflang et al. 2000 for an overview). An implementable model should be:

- Simple (parsimonious). In general, clients require a basic understanding of the logic and benefits of the model. In general, however, clients are no experts in model building, therefore, investigators need models that are easy to communicate. Simple models provide this aspect of implementation;
- Complete on important issues. For a model to be a useful decision-support tool, it has to represent all relevant elements of the problem being studied. In other words, it requires economic coherence. Completeness is relative to the decision problem, organization, and clients. It is extremely difficult to incorporate all relevant problems into one single model without achieving undesirable complexity (Newbold and Bos (1994, p. 563) and Marriott and Tremayne (1988), among others).
- Adaptive. Investigators should be able to quickly adjust the model to changing circumstances under which the MDSS performs. The over-time dynamics, and the associated decision problems may change drastically. Often large-scale models are not very adaptive, therefore, a huge model that works well in one occasion may perform quite poorly in another (Newbold and Bos 1994, p. 552).
- Robust. The model should give economic and statistical logical outcomes. We return to this issue in the model validation section;
- Easy to control and to communicate. The model outcomes should fit the clients' perspectives. An example of a problem that can arise in the communication between investigators and clients is certainty equivalence: the investigators present an *asymmetric* and the clients assume a *symmetric* probability density around the projection of interest. Don (2001) provides an illustrative example. A solution to this problem may be to provide more projections based on clearly defined and distinctive future scenarios.

Note that the investigators cannot fulfill all requirements right from the start. A relatively simple model is often not complete, and a complete model is often not easy to control and communicate. However, if the investigators, experts and clients built up the models in an *evolutionary way*, they can satisfy all requirements to an acceptable level in the end. In such a case, each iteration in the model building process constitutes a small improvement, understandable for the clients. In this way, the model can become more complex in each step, while is still remains easy to communicate. Often, the evolutionary model building process is added to the other aforementioned implementation criteria (Little 1975a, and 1975b).

If we apply the evoluationary model building process to the development of the present study, we obtain an interesting example of the practice of Little's strategies. A report on our experience with this process is given in the Appendix.

[5] Based on Kornelis and Van Meijl (2004).

7. Conclusions

In this study, we have concentrated on the detection of future changes in trends that may affect the dynamics in the agro-food sector, and on the combinaton of opinions of experts in a systematic way. In doing so, we considered three research questions. Our conclusions regarding these questions are as follows:

Which methods exist to elicit influencing factors from group sessions?
Two well-known systematically approaches of expert judgment are the Delphi method, and the cross-impact analysis. Both have their advantages and limitations (see Chapter 2). Another approach is a score card from group sessions, which is used by A&F. The outcome of group sessions with experts are often subject of confounding biases including, among others, groupthink, optimism, anchoring, overreaction, personal involvement, advocacy, and perceived controllability. At the moment, the A&F score-card approach cannot deal with these biases and we see improvements on these issues as relevant topics for future research.

Which methods and concepts exist to combine expert opinion?
Two streams of combination procedures can be identified, viz. mathematical and behavioral approaches. Mathematical procedures can be further classified into axiomatic and Bayesian approaches. On the basis of a literature review, we conclude that:

- the combination of the opinion of multiple experts is an important and non-trivial problem in the use of expert judgment;
- information about experts is often limited, or not available, leading to the use of subjective characteristics;
- the observations of expert opinions often have a hierarchical or multilevel structure;
- the opinion of experts can take the form of a probability statement;
- a natural way of dealing with the combination of expert opinion is a Bayesian data analysis approach;
- the Bayesian data analysis can potentially adopt subjective information about the experts in the form of prior distributions;
- the Bayesian data analysis can be used to estimate multilevel models;
- recently, more and more software packages have become available that perform Bayesian simulation methods;
- the scales on which the opinion of experts is measured is crucial in the selection of appropriate estimation methods.

How to implement methods in such a way that they are accepted and used by stakeholders?
An implementable model should be developed in an evolutionary way, and to optimize its success rate, the model should be: simple (parsimonious), complete on important issues, adaptive, robust, easy to control, and easy to communicate. The developed trendwatch tool of A&F (see Chapter 2) has not fullfilled these criteria yet. For example, it is not complete on important issues. To maximize its success rate it is recommended to make explicit what the ultimate goal is, and what the current limitations are, as this would benefit the communication with the stakeholders. We see these topics as important areas for future research.

Other conclusions
In this study, we investigated the usefulness of multilevel models for the combinations of expert opinions. We used Bayesian data analysis to obtain the parameter estimates. We

applied the advocated approach in two case studies. The results are promising, but the procedures are just a first step into an appropriate combination of expert combination, which has to be complete on important issues, such as the identification of the aforementioned biases. The methodology can be extended into multilevel models with explanatory variables. If we include information on the characteristics of the experts as explanatory variables into the model, it may be an elegant way to deal with interdependencies and/or biases in the subjective opinion of experts. For example, if we adopt 'institute where the expert works' as an explanatory variable in the models, we are able to check if other experts from *other* institutes report different expectations when compared to other experts from the *same* institute. We see this as a promising area for future research

From a modeling point of view, we see an interesting overlap with so-called *meta-analysis* studies, since the latter field of research also makes use of multilevel models. A very important aspect of meta-analyses is the use of a *coding manual*, in which all relevant characteristics about the two levels are collected. In a similar fashion, we may think of a coding manual that collects all relevant characteristics of experts, and which enter the modeling framework as explanatory variables. We see the use as a coding manual as an interesting topic for future research.

References

Armstrong, J.S., 1978, Long-range forecasting, from crystal ball to computer, Wiley and sons, New York.

Berger, James O. 1985, Statistical Decision theory and Bayesian Analysis, (2nd Ed.), Springer-Verlag, New York.

Clemen, Robert T., and Winkler, Robert L. 1999, Combining probability distributions from experts in Risk Analysis, Risk Analysis, 19 (2), 187-203.

Cooke, Roger M. 1991, Experts in uncertainty: opinion and subjective probability in science, Oxford university press, New York.

De Bondt, W. F.M., and R. H. Thaler, 1985, Does the stock market overreact?, Journal of Finance, 40, 793-805.

French, S. 1985, Group consensus probability distributions: a critical survey, In: Bayesian Statistics 2, J. M. Bernardo, M. H. DeGroot, D.V. Lindley, and A.F.M. Smith (Eds.), 183-197, Amsterdam, North-Holland.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 1995, Bayesian data analysis, Chapman & Hall/CRC, Boca Raton.

Genest, C., and Zidek, J.V. 1986, Combining probability distributions: a critical and annotated bibliography, Statistical Science, 1, 114-148.

Greenery, <http://www.thegreenery.nl/>, visited on August 1, 2005

Hardaker, J.B., R.B.M. Huirne and J.R. Anderson, Coping with risk in agriculture. Wallingford, 1997.

Hayes, S.P., jr., 1936, The predictive ability of voters, Journal of Social Psychology, 7, 183-191.

Helmer, O., 1968, Analysis of the future: the Delphi-method, in J. Bright (ed.), Technological forecasting for industry and government, Englewood cliffs, Prentice Haal, N.J.

Helmer, O., 1977, Problems in futures research-Delphi and causal cross-input analysis, Futures, 9.

Janis, I.L., 1972, Victims of groupthink, Boston, Houghton Miflin.

Jaynes, E.T. 1957, Information theory and statistical mechanics, Physics Review, 106, Part I, 620-630; 108, Part II, 171-190.

Koenderink, N.J.J.P., J.L. Top, L.J. van Gool, L.J. van Vliet, "A knowlegde-based method for the automated inspection of natural objects", Submitted to Computer Vision and Image Understanding, July 2004.

Kornelis and Van Meijl, Reducing future uncertainty for decision making processes in agriculture, 2004

Lee, Peter M. 1997, Bayesian Statistics: an introduction, (2nd Ed.), Arnold, London.

LEI, land- en tuinbouwcijfers, Den Haag 2003

Lindley, D.V. 1985, Reconciliation of discrete probability distributions, In: Bayesian Statistics 2, J. M. Bernardo, M. H. DeGroot, D.V. Lindley, and A.F.M. Smith (Eds.), 375-390, Amsterdam, North-Holland.

Linstone, H., and M.Turoff, 1975, The Delphi method: techniques and applications, reading, Addison Wesley, Nass.

Lipscomb, Joseph, Parmigiani, Giovanni, Hasselblad, Vic 1998, Combining expert judgment by hierarchical modeling: an application to physician staffing, Management Science, 44(2), 149-161.

Little, J.D.C., 1970, Models and managers: the concept of a decision calculus, Management Science, 16, B466-B485;

Little, J.D.C., 1975a, BRANDAID: a marketing-mix model, part 1: structure, Operations Research, vol. 23, 628-655;

Little, J.D.C., 1975b, BRANDAID: a marketing-mix model, part 2: implementation, calibration, and case study, Operations Research, vol. 23, 656-673;

Ministerie LNV, <http://www.minlnv.nl/>, visited on August 1, 2004

Myung, In Jae, Ramamoorti, Sridhar, and Baily Jr, Andrew D. 1996, Maximum entropy aggregation of expert predictions, Management Science, 42(10), 1420-1436.

Newbold, P. and T. Bos, 1994, An overview of business forecasting methods. In: Smith, S.L. (Ed.), Introductory business and economic forecasting, 2nd ed., South-Western publishing, Cincinnaty, 550-566.

Productschap Tuinbouw, <http://www.tuinbouw.nl/>, visited on August 1, 2004

WinBugs, 2003, User Manual 1.4, Imperial College & MRC, United Kingdom

Tuinbouw, <http://www.tuinbouw.nl/>, visited on August 1, 2005

Tversky, A., 1974, Assessing uncertainty, Journal of the Royal Statistical Society, Series B, 36, 148-159.

Tyebjee, T.T., 1987, Behavioral biases in new product forecasting, International Journal of Forecasting, 3, 393-404.

Van Klink, A., Visser, E.J., “Innovation in Dutch horticulture: fresh ideas in fresh logistics”, Tijdschrift voor Economische en Sociale Geografie, vol. 95, no.3, pp340-346, 2004.

Wright, G., P. Ayton, 1989, Decision time, subjective-probability, and task-difficulty, Memory and Cognition, 16, 2,176-185.

Ying, Yun, Lynn Kuo and Gim S. Seow, “Forecasting Stock Prices Using a hierarchical Bayesian Approach”. In: Journal of forecasting. 24 (2005), 39-59.

Appendix

The development over time of the present study gives an illustrative example of the intuition behind the model building process that was developed by Little (1970, 1975a, 1975b).
The ultimate goal (and ideal point) of the research was to come up with an easy-to-implement tool for expert combination, which would deal with all possible biases in the (subjective) judgments of experts. We started *simple*. In our case, 'simple' implied the use of techniques that matched the skills of the assessors, would provide better information than an unweighted mean, and, would be an appropriate step into the direction of the ultimate goal (often, ad hoc procedures may work very well in the short-run, but turn out to be a "waste of time" in the long-run, as they do not bring the researchers closer to the end goal). The obvious choice for us, was to use maximum entropy techniques, because, at WUR, many assessors have sufficient knowledge about this technique to use it in a proper way. Hence, we applied the study of Myung et al. (1996). This approach had some major limitations (see Chapter 2). Trying to "solve" these "problems", we became aware of the fact that data on expert opinion has a hierarchical structure. An appropriate analysis of hierarchical data requires hierarchical models. This lead to the proposed multilevel approach (see Chapter 4). Using only hierarchical models for continuous scales lead to limitations. So, we studied literature on to use multilevel models concerning categorical data (see chapter 4 and 5).
In sum, each aforementioned step lead to new developments and to more complex models, but, because the limitations of previous procedures were made explicit in each step, the adopted procedures were still easy to communicate and understand. Notice, that the ultimate goal (the combination of expert opinion) remains the same in each step, but that the way leading to it, does not appear to be a straight line. It is rather a meandering series of steps. And each following step brings us closer to our ultimate goal: combining expert opinion in such a way that the confounding influences of biases are minimized.