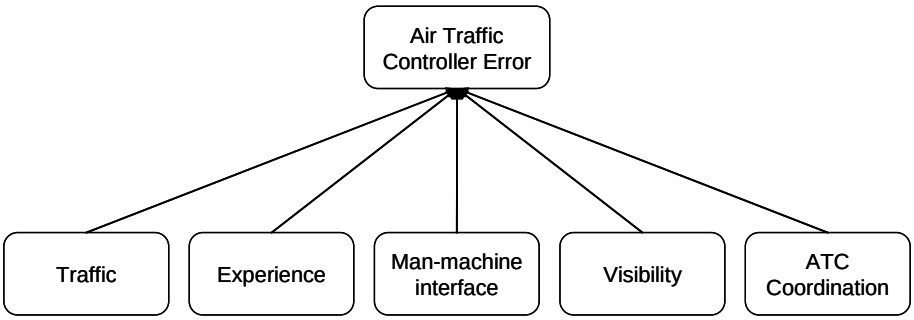
Nationaal Lucht- en Ruimtevaartlaboratorium
National Aerospace Laboratory NLR

Executive summary

A generic air traffic controller performance model for application in a causal model of air transport



Problem area
The Netherlands Ministry of Transport has initiated a research effort to develop a causal model for aviation safety. The purpose of the model is to describe the air traffic system and its safety functions in such a way that it is possible to analyse risk reduction alternatives and to serve as a means of communication between experts and managers within the industry. The model combines Event Sequence Diagrams, Fault Trees and Bayesian Belief Nets (BBNs). In order to account for the influence of the human operator on accident causation, that role must be properly represented in the causal model. The human operator is best represented in a Bayesian Belief Net because it involves primarily 'soft' influences rather than deterministic cause-effect relations. The objective of this part of the study is to develop a quantified model for air traffic controller performance for a causal model of air transport safety.

Description of work
The development of a quantified model of air traffic controller performance is based on existing research on human performance, compatibility with other parts of the causal model, and the mathematical specifics of Bayesian Belief Nets. The BBNs in the causal model use continuous variables rather than 'conventional' BBNs that use discrete variables. The variables are represented by one dimensional marginal distributions and the relations between the variables are represented by rank correlations.

Report no.
NLR-CR-2007-593

Author(s)
A.L.C. Roelen
G.B. van Baren
P.H. Lin
O. Morales
D. Kurowicka
R.M. Cooke

Report classification
UNCLASSIFIED

Date
June 2008

Knowledge area(s)
Safety & Security

Descriptor(s)
Safety
Risk assessment
Human reliability

Dependencies are quantified by means of expert judgement; the information was obtained according to the protocol developed by Cooke & Goossens [2000].

Results and conclusions

Two type of model have been developed: the Radar Controller model is used to represent both the area controller and the approach controller, the Tower Controller model is used to represent the runway controller and the ground controller. The two models are similar in structure but differ in the quantification of the dependencies. This report only describes the dependency information for the Radar Controller model.

The focus of this study was development of a practicable and working model, ready for use within the causal model. Consequently the model is relatively simple, and some of the influences might require further refinement at a later stage.

It is recommended to conduct a separate expert elicitation session to quantify the dependencies in the Tower Controller model.

Nationaal Lucht- en Ruimtevaartlaboratorium, National Aerospace Laboratory NLR

Anthony Fokkerweg 2, 1059 CM Amsterdam,
P.O. Box 90502, 1006 BM Amsterdam, The Netherlands
Telephone +31 20 511 31 13, Fax +31 20 511 32 10, Web site: www.nlr.nl

Nationaal Lucht- en Ruimtevaartlaboratorium
National Aerospace Laboratory NLR

NLR-CR-2007-593

A generic air traffic controller performance model for application in a causal model of air transport

A.L.C. Roelen, G.B. van Baren, P.H. Lin[1], O. Morales[1], D. Kurowicka[1] and R.M. Cooke[1]

[1] TU Delft

No part of this report may be reproduced and/or disclosed, in any form or by any means without the prior written permission of the owner.

Customer	Ministry of Transport, Public Works and Water Management
Contract number	Ra-05.041 / DGL 5.50.2.4019
Owner	Ministry of Transport, Public Works and Water Management
Division	Air Transport
Distribution	Limited
Classification of title	Unclassified
	June 2008

Approved by:

Author	Reviewer	Managing department
13/06/08	13/6/08	13/6/8

Contents

1 Introduction **7**

1.1 Background 7

1.2 Objective and requirements 8

1.3 Research approach 8

1.4 Ackowledgements 8

2 Review of existing human performance models **9**

3 Linking with other parts of the causal risk model **11**

4 Mathematics of Bayesian Belief Nets **13**

5 Model development **14**

5.1 General description of air traffic control 14

5.2 Technology 15

5.3 ATCo performance shaping factors 17

5.4 Model structure 19

6 Model quantification **20**

6.1.1 Quantification of the marginal distributions 20

6.1.2 Experience 21

6.1.3 Traffic 21

6.1.4 Man Machine Interface 22

6.1.5 ATC communication / co-ordination 23

6.1.6 Visibility procedure 24

6.2 Expert elicitation 25

7 Quantified model **26**

8 Conclusions and recommendations **27**

References **28**

Appendix A	**List of experts**	**31**
Appendix B	**Elicitation protocol**	**32**
Appendix C	**Answers to calibration questions.**	**44**
Appendix D	**Expert's results for probability distributions**	**46**
Appendix E	**Description of the expert elicitation results**	**47**

Abbreviations

ACARS	Aircraft Communication Addressing and reporting System
ACC	Area Control Centre
ANSP	Air Navigation Service Provider
APP	Approach
A-SMGCS	Advanced Surface Movement Guidance and Control System
ATC	Air Traffic Control
ATCo	Air Traffic Controller
ATM	Air Traffic Management
BZO	Bijzonder Zicht Operatie (limited visibility operation)
BBN	Bayesian Belief Net
CRDA	Converging Runway Display Aid
DM	Decision maker
ESD	Event Sequence Diagram
EWDM	Equal Weight Decision Maker
FAST	Final Approach Spacing Tool
GND	Ground
GWDM	Global Weight Decision Maker
IWDM	Item Weight Decision Maker
MSAW	Minimum Safe Altitude Warning
MUAC	Maastricht Upper Area Control Centre
PSF	Performance Shaping Factor
RIAS	Runway Incursion Alerting System
RVR	Runway Visual Range
STCA	Short Term Conflict Alert
TMA	Terminal Manoeuvring Area
TWR	Tower
VHF	Very High Frequency

This page is intentionally left blank.

1 Introduction

1.1 Background

The Netherlands Ministry of Transport has initiated a research effort to develop a causal model for aviation safety [Ale et al 2005]. The purpose of the model is to describe the air traffic system and its safety functions in such a way that it is possible to analyze risk reduction alternatives and that it will serve as a means of communication between experts and managers within the industry. The model combines Event Sequence Diagrams, Fault Trees and Bayesian Belief Nets into a single structure. The US Federal Aviation Administration (FAA) is developing a similar integrated risk model in support of the Systems Approach to Safety Oversight (SASO) program. The proposed risk model architecture also introduces a hybrid causal model of Event Sequence Diagrams, Fault Trees and Bayesian Belief Networks. Because of the similarities between the two research efforts it was decided to frequently exchange results and if possible use similar components in the model and the model structure. Both causal models use the same backbone structure of generic accident scenarios. The ESD methodology is used for representing accident scenarios. Identified accident scenarios are e.g. abrupt manoeuvre, uncontrolled collision with ground, controlled flight into terrain, forced landing, mid-air collision, collision on ground, structure overload, and fire/explosion. In this way, 33 generic accident scenarios have been developed. An ESD consists of an initiating event, pivotal event(s), and an end state. Where necessary, the initiating and pivotal events are detailed in a submodel which can be a Fault Tree or a Bayesian Belief Net. Initiating and pivotal events may involve a human operator. The human operator plays an essential role at the operational level of any risk bearing activity. In order to account for the influence of the human operator on accident causation, that role must be properly represented in the causal risk model. The human operator is best represented in a Bayesian Belief Net because it involves primarily ‘soft’ influences rather than deterministic cause-effect relations. More details on the development of the causal model for air transport safety can be found in interim progress reports [Ale 2007]. It is envisioned that three types of human operators will be explicitly represented in the causal model for air transport safety:

- Flight crew,
- Air traffic controller,
- Maintenance technician.

The development of a generic flight crew performance model has been reported in Roelen et al [2007]. The underlying report describes the development of an air traffic controller performance model.

1.2 Objective and requirements

The objective of this part of the study is to develop a quantified model for air traffic controller performance for a causal risk model of air transport safety. The purpose of the human performance model is to describe and quantify the influences on the probability of human failure in the fault trees describing initiating and pivotal events that constitute the ESDs of the model. The ESDs represent generic accident scenarios. A generic air traffic controller performance model is required that can be used to represent the role of the air traffic controller for different situations. The human performance model should focus on those human actions that influence the accident scenarios. The model should remain as simple as possible, focusing on the most important influential factors only. These factors should have a clear and unambiguous definition and rating scale. This is essential as the quantification of the model should be based either on available data or on expert opinion. The model will be a Bayesian Belief Net to be able to represent soft influences rather than deterministic cause-effect relations. Finally, the model should allow representation of managerial and organisational influences on air traffic controller performance.

1.3 Research approach

The development of a quantified model of air traffic controller performance is based on existing research on human performance, the requirement to be able to link-up with other parts of the causal risk model, and the mathematical specifics of Bayesian Belief Nets. Each of these topics has its own characteristics and requirements, which we tried to merge as much as possible into a practicable model. A similar approach was followed in the development of a generic flight crew performance model [Roelen et al 2007]. The resulting qualitative model was quantified with a combination of existing data and expert judgement. Consulted experts were Air Traffic Controllers from the Maastricht Upper Area Control Centre (MUAC). Quantification of the model concerns quantification of the marginal distributions of the nodes in the model as well as of dependencies between the nodes.

1.4 Ackowledgements

The authors are indebted to Keith Cartmale from the Maastricht Upper Area Control Centre (MUAC) for his help in organising the expert elicitation meeting, and to the experts that kindly participated in the elicitation process.

2 Review of existing human performance models

Human error modelling started in earnest in the 1970s. The vast majority of the early development of techniques for hazard identification and hazard representation was based on modelling hardware failures. Since human behaviour and failure is different than hardware performance and failure, new techniques had to be developed to incorporate it. In particular human behaviour is strongly influenced by recovery actions as well as initial failures. A classical approach in human error quantification is to use basic error probabilities that are modified to account for specific circumstances or contexts. Human error probabilities for general types of tasks are adjusted for the influence of possible circumstances or contexts by the application of performance shaping factors [Swain & Guttmann 1983]:

$$P(error) = HEP * \sum_{i=1}^{N} PSF_i * W_i \qquad (1)$$

HEP is the human error probability for a certain generic task, PSF_i is the i-th performance shaping factor and W_i is the weight of PSF_i. Examples of generic tasks are selecting a control, operating a control, etc. Examples of performance shaping factors are workload, time since last training, etc. Performing shaping factors are normally not seen to be causal in a strict sense, but are generally thought of as factors that contribute to how erroneous actions manifest themselves [Hollnagel 1998]. A strict methodology for deriving performance shaping factors does not exist. They are known, through experience, to have consequences for how the task is carried out and how frequent errors occur. Two of the most often used techniques that follow this approach are HEART (Human Error Assessment and Reduction Technique) [Kirwan 1994] and THERP (Technique for Human Error Rate Prediction) [Swain & Guttmann 1983]. Application of these and similar first generation methods usually require a task analysis in combination with engineering techniques such as fault and event tree logic to identify where errors are important. These latter methods are better at dealing with omissions/procedures and are not good at supporting the identification of errors of commission. Errors of commission are particularly difficult to identify with techniques that only look at the logical operation and failures of the system according to design and procedures. Identifying such errors requires special support that is not readily available in the traditional way that fault trees and event trees were build to include human error [Hale et al 1999]. In particular the construction of fault trees from planned/intended system performance, by systematically incorporating the omission or failure of each intended step, will incorporate many slips and lapses, but miss many mistakes and some violations. A second drawback of the way in which first generation methods combine human error probabilities and performance shaping factors is the assumption that the effects of the

context are independent. This is not a very reasonable assumption. As an example, stress cannot be considered independent of workload. Second generation methods have been developed that try to cope with these issues. Examples of second generation methods are ATHEANA (A Technique for Human Event Analysis) [NUREG 2000] and CREAM (Cognitive Reliability and Analysis Method). The CREAM method starts with identifying the context and common performance conditions [Hollnagel 1998]. The generic task is then associated with a set of context specific error modes.

Some of the limitations of first and second generation human methods can be overcome by using BBNs to represent the influence of the context or performance shaping factors on error probability. Dependencies and interactions between performance shaping factors are easily represented in a BBN structure. In such a BBN, overall error probability is calculated as a conditional probability of a particular set of performance shaping factors rather than as a sum of separate and independent performance shaping factors. It is proposed here to represent the air traffic controller performance model as a Bayesian Belief Net with air traffic controller error as the child node and various performance shaping factors as the parent nodes, see Figure 1.

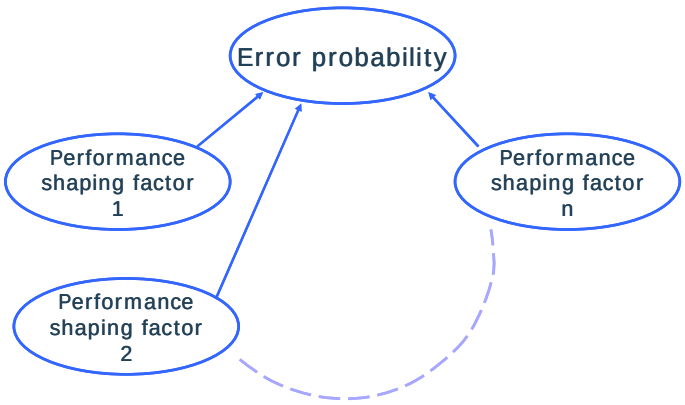


Figure 1: Human performance model schematic

3 Linking with other parts of the causal risk model

The design of the causal risk model for air transport calls for a combination of three modelling techniques in a single model: Event Sequence Diagrams (ESDs), Fault Trees (FTs) and Bayesian Belief Nets (BBNs), see Figure 2 [Ale et al 2007]. The ESDs represent generic accident scenarios at the most abstract level of the model [Roelen & Wever 2005]. Fault Trees link to the initiating events and pivotal events of the ESDs and describe them in a more detailed manner as a sequence of barrier failures. These base events of the fault trees include events representing human reliability, such as for instance 'ATCo failure to respond to MSAW'. These base events involving human reliability are detailed further as Bayesian Belief Nets.

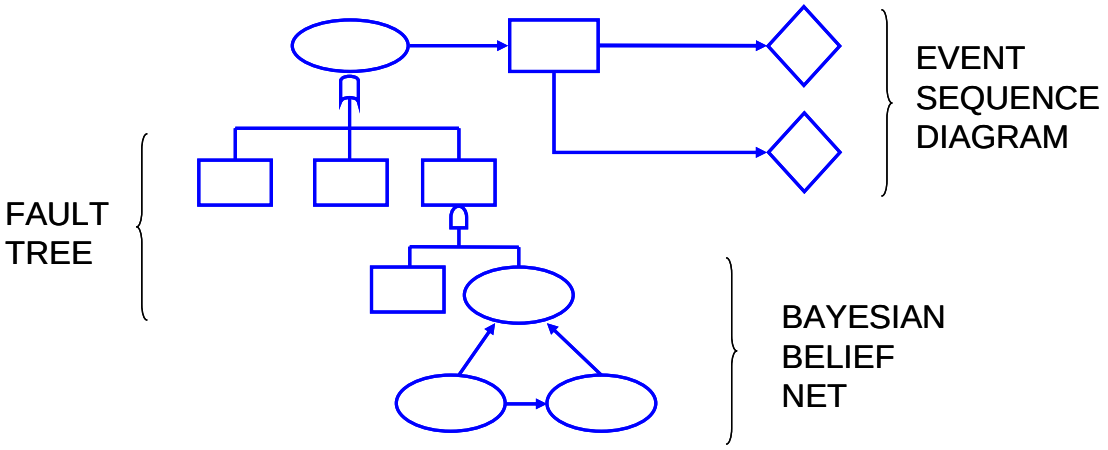


Figure 2: Overall construction of the causal risk model as a combination of ESDs, FTs and BBNs[1]

The actual air traffic controller error probabilities follow from the fault trees. We propose to use the associated event in the fault tree to estimate the basic human error probability instead of deriving this from a task analysis as is the case in first generation human performance methods. Mapping of accident and incident data on those fault trees will ensure that both errors of omission as well as errors of commission will be captured in the model, because the fault trees contain both types of errors.

The air traffic controller performance model must be able to represent managerial and organisational influences on human performance. In the causal risk model, safety management is described as providing the resources and criteria for the frontline workforce to operate safely. Resources and criteria are generically described by delivery systems for both hardware and

[1] In this schematic drawing, only the initiating event of the ESD is described by fault trees and BBNs. In the causal risk model however, the initiating as well as the pivotal events of the ESD are described by fault trees and BBNs.

human elements. The human part is governed by the following delivery systems [Bellamy et al 1999]

- Procedures, output goals and plans
- Technology interface
- Availability of personnel
- Competence
- Communication and coordination
- Commitment.

In order to represent managerial influences on air traffic controller performance, it is proposed to describe the influences on air traffic controller performance in terms of the delivery systems, see Figure 3.

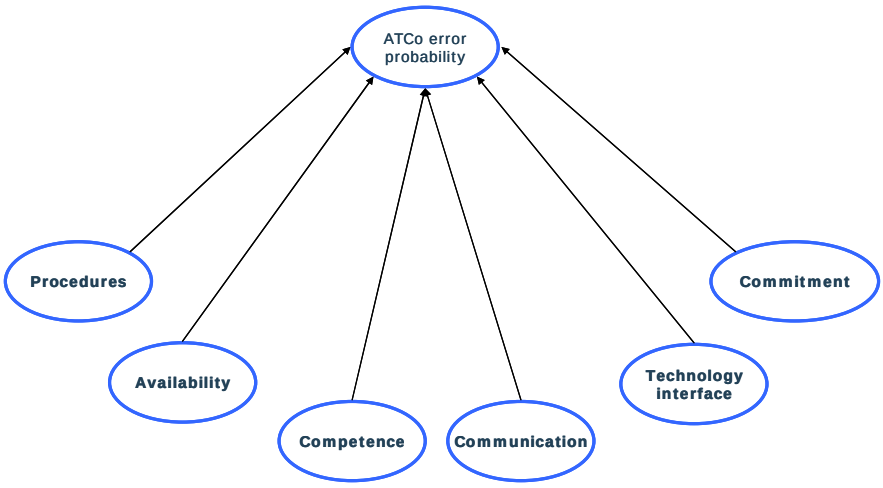


Figure 3: Linking managerial influences to the air traffic controller error probability

The delivery systems are too generic to act directly as performance shaping factors, but they provide a framework for PSF selection. The aim is to cover all dimensions of safety management by selecting the PSFs in such a way that each of the six delivery systems is represented.

4 Mathematics of Bayesian Belief Nets

A detailed description of the mathematics of BBNs is provided in Morales et al [2007]. In the underlying report, several elements that are relevant for the choices made in the development of the model are repeated. The Bayesian Belief Nets that are used in the causal risk model are BBNs with continuous variables rather than 'conventional' BBNs that use discrete variables [Ale et al 2007]. When the variables (the nodes in the graph) would be represented as discrete variables and all conditional probabilities would be quantified in tables, the assessment burden can become enormous. When e.g. a child node has 7 parent nodes and each node is discretised to take 3 possible values, a conditional probability table consisting of 3·(37)=6561 entries has to be created, which means that 6561 conditional probabilities have to be assessed. Representation of the variables as continuous variables implied until recently that the distribution of the variables was restricted to a joint normal distribution. Hanea et al [2006] introduced the 'copula-vine' approach. In this approach, the variables are represented by one dimensional marginal distributions (not necessarily normal) and the relations between the variables are represented by rank correlations. The number of rank correlations to be assessed equals the number of arcs in the model. No assumption has to be made on the joint distribution; instead a copula is chosen. A copula is a multivariate cumulative distribution function defined on the n-dimensional unit cube $[0, 1]^n$ such that every marginal distribution is uniform on the interval [0, 1] [Bedford & Cooke 2002]. The copula contains all of the information on the nature of the dependence between the random variables that can be given without knowing the marginal distributions, but gives no information on the marginal distributions. In effect the information on the marginal distribution and the information on the dependence are neatly separated from each other. The joint distribution of the BBN is stipulated by Monte Carlo simulation.

Requirements and assumptions are [Morales et al 2007]:

- the copula represents (conditional) independence as zero (conditional) correlation.
- the rank correlations between the variables are constant; i.e. a situation where X and Y are positively correlated when variable Z takes low values, but are negatively correlated when Z takes high values cannot be represented.
- variables have to be monotonic.

Dependencies are quantified by existing data or by means of expert judgement. When expert judgement is involved, elicitation is done according to the protocol developed by Cooke and Goossens [2000]. The information on the dependencies between certain variables is elicited via the rank and conditional rank correlations, i.e. we first ask a ranking of the variables from the largest to the smallest influence. For the 'most influential' variable we then ask for the

probability that the variable 'ATCo error' is above its median value given that the most influential variable is above its median value. Finally, the experts are asked to estimate the other variables' influence as a fraction of the 'most influential' variable's influence. This information together with the marginal distributions for each node is enough to compute the (un)conditional rank correlations required by the model. The procedure is explained in detail in Appendix E.

A consequence of the use of expert judgement is that the variables under consideration must be expressed in objectively quantifiable units. 'Objectively' denotes that a particular value of that unit has the same meaning for expert A and expert B. This requirement forces us to be very careful in selecting and defining the variables that are included in the model. As an example, consider the variable 'safety culture'. We might be tempted to include this as a variable in the model and apply a 5-point rating scale: very poor, poor, medium, good, very good. We could then ask experts to quantify the effect of safety culture on error probability. The problem here is that as long as we have not strictly defined what we mean with poor safety culture, every expert will interpret it differently and results from different experts cannot be combined.

5 Model development

5.1 General description of air traffic control

The air traffic controller (ATCo) is the front line operator in Air Traffic Management (ATM) whose primary responsibility is the safe separation between aircraft. To meet this responsibility, the ATCo issues instructions via VHF radio that need to be obeyed by the flight crew, unless they are of the opinion that following the controller's instruction compromises flight safety. Communication involves instructions according to a protocol that is laid down in ICAO Annex 10, part II, Chapter 5: R/T. Air-ground radiotelephony must be conducted in the language normally used by the station on the ground or, preferably, in the English language.

For practical reasons, the airspace is divided into different sectors which each normally have a separate ATCo. Control of a particular aircraft is handed over from one ATCo to another when the aircraft crosses a sector boundary. Different types of ATCos exist to properly handle the various phases of flight. The exact organisation will depend on the specific situation, subject to such characteristics as the complexity of the airport, the number of runways and the amount of traffic, but generally speaking the following types of controller are distinguished:

The *ground controller* supports aircraft on the ground during departure from the gate up to the holding point just before the take-off runway, and during arrival from the moment the aircraft clears the runway after landing up to parking at the gate. The *tower controller* takes over from

the ground controller for departing aircraft as they enter the runway and supervises the aircraft until they have reached 2000 ft altitude at which point control is handed over to the departure controller. For arriving aircraft, the tower controller brings the aircraft in to land from the point during the approach when they are lined-up with the runway and the ILS is established up to the point where the aircraft clears the runway.

The *departure controller* takes over departing aircraft from the tower controller up to the point where they are free of all conflict with other aircraft at and around the airport. At this point the *area controller* takes over for the transit to the following sector/area. When the aircraft is within 50 to 30 nautical miles from the arrival airport the aircraft is handed over to the *approach controller* who takes the aircraft from there until it is lined up for the runway and ILS is established and the tower controller takes over.

The ground controller and tower controller are located in the traffic control tower and predominantly monitor the traffic by looking outside. Departure controllers, area controllers and approach controllers monitor traffic via radar screens and are usually not located in the control tower.

5.2 Technology

Air traffic controllers are supported by various technological systems to perform their functions efficiently. The following systems are considered most relevant:

- Primary and Secondary Radar
- Ground radar, possibly with ground labels
- Runway Incursion Alerting System (RIAS)
- Converging Runway Display Aid (CRDA)
- Short Term Conflict Alert (STCA)
- Minimum Safe Altitude Warning (MSAW)
- Final Approach Spacing Tool (FAST)
- Flight progress strips

<u>Primary and Secondary Radar</u>

A controller's radar screen depicts the various aircraft flying within his area of responsibility. Primary radar sends out an energy pulse of which a small part is reflected by objects such as aircraft. A controller sees these objects as 'blips' on his radar screen. Secondary radar interrogates a transponder on-board the aircraft, upon which the transponder sends back relevant information. Each aircraft on the controller's radar screen is indicated by a label that contains specific information like the flight number and altitude of the aircraft.

Ground radar

Ground radar is used to depict the position of aircraft on the airport to help prevent ground collisions. Most advanced ground radar systems also depict labels with aircraft specific information.

Runway Incurion Alerting System (RIAS)

RIAS is a protection system against runway incursions. It will give an alert to the runway controller when a runway incursion occurs.

Converging Runway Display Aid

CRDA is used on some airports that allow simultaneous approaches on two converging runways. The tool is intended to prevent mid-air collisions in case of simultaneous missed approaches by providing a certain stagger between the arriving aircraft. The size of the stagger depends on the aircraft speed and performance variations and the runway geometries. The CRDA converts the geometry of the converging approaches into a single runway approach geometry. It projects an electronic ghost image of an aircraft approaching to one runway on the approach path of the other converging runway allowing the approach controller to accurately monitor the required stagger distances [Abbink 1996].

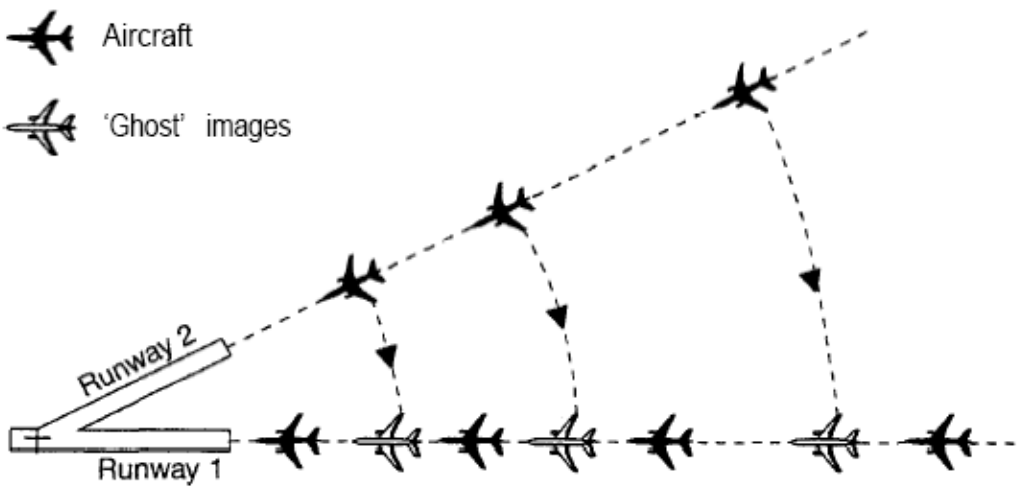


Figure 4: Converging Runway Display Aid schematic

Short Term Conflict Alert

STCA is a warning system designed to help prevent mid-air collisions The system detects potential conflicts between aircraft by on projecting aircraft tracking data forward in time and gives air traffic controllers an automated warning in case of a conflict.

Minimum Safe Altitude Warning

MSAW provides a visual and aural alert to a controller whenever an aircraft descends below, or is predicted by the software to descend below, a predetermined safe altitude[2].

FAST

FAST assist the approach controller to obtain an optimal line -up of arriving aircraft by calculating optimal vectoring advisories and presenting them on a display to the controller.

Flight progress strips

Flight progress strips are paper strips containing flight information. The strips are used by the air traffic controllers to keep track of all aircraft that are under his control. Handing over of an aircraft from one controller to another is accompanied by a physical handover of the flight progress strip. Any operational change relative to the original flight plan like a change in departure time, is easily tracked by the controller by simply making notes on the strip. Some ANSPs are in the process of switching over from paper strips to electronic flight progress strips.

5.3 ATCo performance shaping factors

An initial set of ATCo performance shaping factors has been selected after a review of literature [Kirwan 2001, Shorrock & Kirwan 2002, Shorrock 2005, Shorrock 2006, Rognin & Blanquart 2001, Van Es 2004, Subotic et al 2006] and preliminary analysis of a large sample of accidents and incidents[3]. Table 1 presents the initial list of selected performance shaping factors together with the associated delivery systems. An appropriate performance shaping factor for the delivery system 'commitment' could not be defined because it was not possible to express the level of commitment in an objectively quantifiable unit.

Table 1: Initial selection of performance shaping factors.

Performance shaping factor	Delivery system
Workload	Availability
Experience	Competence
Man-machine interface	Technology interface
Visibility	Procedures
ATC coordination	Communication and coordination

Each of these factors in discussed in detail in the following sections.

[2] Under the condition that the aircraft is equipped with an altitude encoding transponder (Mode C).

[3] Narratives from ECCAIRS, 1990-2005, commercial operations, aircraft heavier than 5,700 kg MTOW.

Workload, expressed as Traffic

Workload for the controller is expressed in this model as the number of aircraft simultaneously under control of a single controller. Other factors such as traffic diversity and the complexity of air routes are known to influence controller workload as well, but as of yet we have not been able to represent these characteristics in objectively quantifiable units.

Experience

Experience of an air traffic controller can be expressed as the overall experience as an air traffic controller and as experience on a particular position, like approach controller. In this study we refer to experience as the number of years working as an air traffic controller in the current position because this is considered to be the most relevant.

Man-machine interface

Four different states are distinguished, based on the availability of supporting systems:

1. VHF radio only,
2. VHF radio and primary radar,
3. VHF radio, primary radar and secondary radar,
4. VHF radio, primary radar, secondary radar and additional tools like RIASS, STCA, etc.

These classes were defined rather arbitrarily. The aim is to be able to describe all possible situations that exist in different regions of the world, without going into much detail.

Visibility

Activities of air traffic controllers in the control tower are directly influenced by the visibility conditions. When visibility is reduced the controllers are less able to observe directly what is happening on the airfield and the working procedures will change when visibility drops below certain pre-set threshold values. At Schiphol the following visibility conditions are distinguished:

- Normal operations: RVR above 1500 m and cloud base above 300 ft.
- BZO A: RVR between 550 m and 1500 m or cloud base between 200 ft and 300 ft (associated with Cat 1 approaches).
- BZO B: RVR between 350 m and 550 m or cloud base below 200 ft (associated with Cat II approaches).
- BZO C: RVR between 200 m and 350 m (associated with Cat 3a approaches).
- BZO D: RVR below 200 m (associated with Cat 3b approaches).

ATC coordination

Air traffic controllers need to coordinate their activities with those of other controllers in their own organisation as well as controllers in other organisations, e.g. when an aircraft is handed over from one sector to another. In this model we distinguish between situations where two coordinating controllers are located in the same room and can communicate directly without technology aids and situations in which two coordinating controllers are not located in the same room and communication relies on telephone, radio, fax, etc.

5.4 Model structure

The generic model structure is presented in Figure 5. The probability of 'Air Traffic Controller error' is influenced by the variables 'traffic', 'experience', 'man-machine interface', 'visibility' and 'ATC coordination' as defined in section 5.3. The parent nodes are assumed to be independent. As was explained in the previous section there are different types of controllers. For the purpose of the model, we distinguish two types: the Radar Controller and the Tower Controller. The Radar Controller model is used to represent both the area controller and the approach controller, the Tower Controller model is used to represent the runway controller and the ground controller. Although the tasks of for instance the area controller and the approach controller are different, we assume that the influences on their performance can be described by the same model. The advantage of this assumption is the reduction in data requirements, which is particularly important because we need to quantify the models with the help of expert judgement and air traffic controllers are notoriously difficult to get hold of.

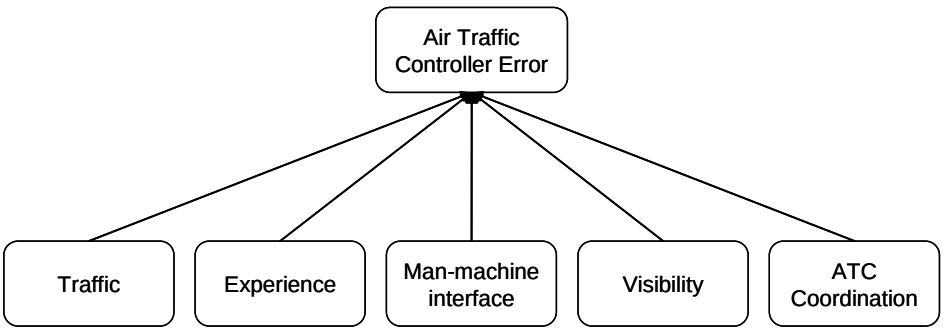


Figure 5: Generic air traffic controller performance model structure

6 Model quantification

Quantification of the model concerns quantification of the marginal distributions of the nodes in the model as well as of the dependencies between the nodes.

6.1.1 Quantification of the marginal distributions

Table 2 lists the nodes, their definition and how the marginal distributions were derived.

Table 2: ATCo model nodes and their definitions

Node	Definition	Unit	Source for marginal distribution
Experience	Number of years working as an ATCo at current position	Number of years	Eurocontrol report on Age, Experience and Automation in European ATC – Survey in the ECAC area [Eurocontrol 2004]
Traffic	Number of aircraft simultaneously under control by ATCo	Number of aircraft	CFMU data
Man Machine Interface (MMI)	Classification of the interface that is used by the ATCo. Four classes are distinguished, based on the availability of supporting systems: 1. radio only 2. radio and primary radar 3. radio and primary and secondary radar 4. radio and primary and secondary radar and additional tools (e.g. RIASS, STCA)	1 – 4	General information on European airports [Eurocontrol 2005] and [Eurocontrol 2006].
ATC co-ordination	Whether or not the communication/ co-ordination between ATCos takes place in the same room. In the latter case it is assumed that telephone or radio is used.	yes/no	Actual situation at Schiphol.
Visibility procedure	Visibility procedure in use, subdivided into 5 classes: 1. normal operations: RVR above 1500m and cloud base above 300ft. 2. BZO A: RVR between 550m and 1500m or cloud base between 200ft and 300ft (associated with Cat I) 3. BZO B: RVR between 350m and 550m or cloud base below 200ft (associated with Cat II) 4. BZO C: RVR between 200m and 350m (associated with Cat 3A) 5. BZO D: RVR below 200m (associated with Cat 3B) BZO A – D are in place at Schiphol airport. BZO stands for 'Bijzondere Zicht Operatie' (Limited Visibility Operations). RVR stands for Runway Visual Range.	1 – 5	Schiphol operational information in combination with meteorological data

The quantification of the marginal distributions is discussed for each node separately in the following subsections.

6.1.2 Experience

Data on experience of ATCos are obtained from a Eurocontrol survey on age, experience, and automation in European ATC in the ECAC area [Eurocontrol 2004]. The distribution for the experience on the current positions is presented in Figure 6. 1079 ATCos from 28 countries participated in the survey. The average age of the subjects was 38.17 years, with a standard deviation of 8.94. The youngest controller in the sample was 21 years old at the time of the survey, the oldest controller was 60. The controllers questioned have a job experience on their current position of 121 months on average. Standard deviation is 97.16 months, the maximum is 408 months and the minimum zero.

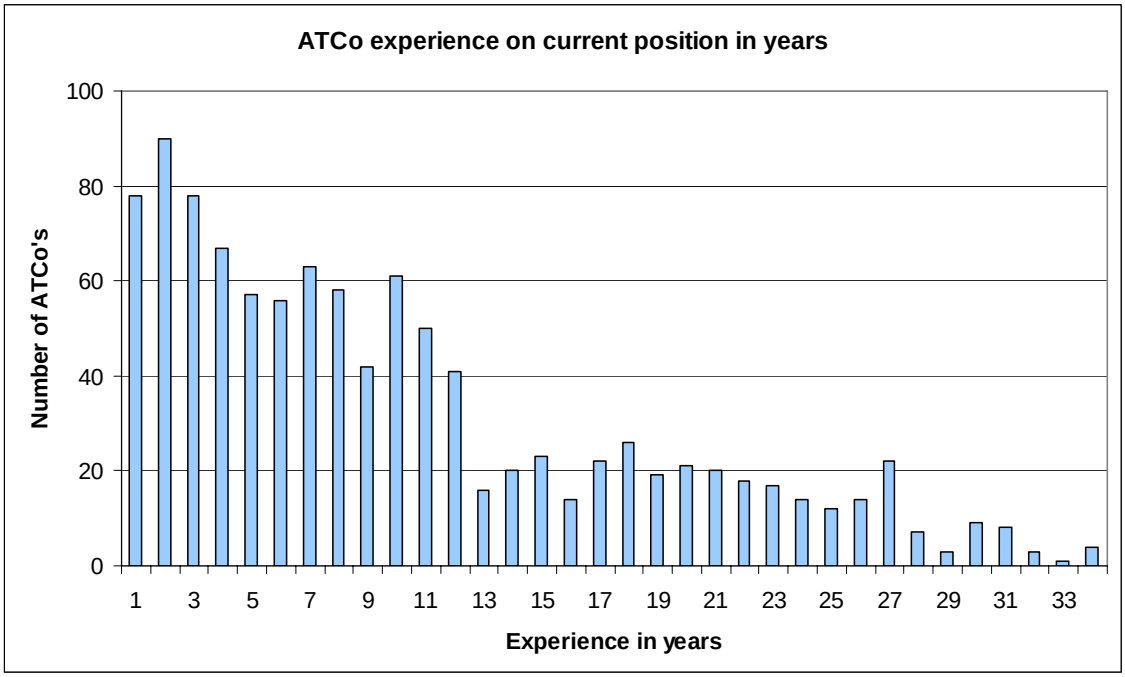


Figure 6: Histogram of ATCo experience

In the model, we assume that the distribution of experience is independent of the type of air traffic control (tower or radar).

6.1.3 Traffic

The distribution of traffic, defined as the number of aircraft simultaneously under control of a single ATCo, is based on Central Flow Management Unit (CFMU) data of four days[4] of movements into the Schiphol ACC sector via entry ARTIP. The data lists the number of aircraft that entered the sector per time frame of one hour. Using estimates for the size of the airspace (100 km cross section) and the average speed of the aircraft (350 kts), the average time of an

[4] 6, 7, 10 and 14 September 2004

aircraft in the sector as well as the number of aircraft simultaneously in the sector can be calculated. The resulting distribution is shown in Figure 7. This distribution is assumed to be independent of the type of controller (tower or radar).

Figure 7: Number of aircraft under control simultaneously

6.1.4 Man Machine Interface

For classes of man-machine interface are distinguished in the model:

1. radio only,
2. radio and primary radar,
3. radio and primary and secondary radar, and
4. radio and primary and secondary radar and additional tools (e.g. A-SMGCS, STCA).

In Eurocontrol's 2005 status report [Eurocontrol 2006] on the European Convergence and Implementation Plan (ECIP), it is shown that about 80% of aircraft movements within the ECAC ACC's are supported by STCA. This number is taken as a representative for class 4 probability for ACC and APP controllers. According to Eurocontrol's Surveillance Strategy for ECAC [Eurocontrol 2005], Secondary Surveillance Radar (SSR) is widely implemented throughout the ECAC area. As 'widely' is not further specified, we (arbitrarily) assume that this refers to 99% of the ECAC area. If we further assume that class 1 'radio only' only can occur due to a radar failure and that such failure occurs with a probability 1 in 10,000, we obtain the following probabilities for the four classes:

Table 3: Man machine interface distribution for radar controller

Class	Probability of occurrence for ACC and APP
1. Radio only	0.0001
2. Radio and primary radar	0.0099
3. Radio and primary and secondary radar	0.1900
4. Radio and primary and secondary radar and additional tools (e.g. A-SMGCS, STCA).	0.8000

For TWR and GND controllers it is more appropriate to look at implementation of A-SMGCS (Advanced Surface Movement Guidance & Control System) Level 1. According to Eurocontrol [2005] implementation is fully or partially completed in 21% of the ECAC states. This number is considered representative for the class 4 probability for TWR and GND controllers. According to Eurocontrol's Surveillance Strategy for ECAC [Eurocontrol 2005], Secondary Surveillance Radar (SSR) is widely implemented throughout the ECAC area. As 'widely' is not further specified, we (arbitrarily) assume that this refers to 99% of the ECAC area. If we further assume that class 1 'radio only' only can occur due to a radar failure and that such failure occurs with a probability 1 out of 10,000 movements, we obtain the following probabilities for the four classes:

Table 4: Man machine interface distribution for Tower Controller

Class	Probability of occurrence for TWR and GND
1. Radio only	0.0001
2. Radio and primary radar	0.0099
3. Radio and primary and secondary radar	0.7800
4. Radio and primary and secondary radar and additional tools (e.g. A-SMGCS, STCA).	0.2100

6.1.5 ATC communication / co-ordination

For the communication or co-ordination between ATCos a distinction is made between the communication / co-ordination taking place in the same room or not. In the latter case it is assumed that telephone or radio is used. From the situation at Schiphol airport the following information is known:

- GND ATCo usually communicates with one other GND ATCo and two TWR ATCos (e.g. two runway controllers).
- TWR ATCo communicates with one GND, one other TWR, and one APP controller.
- APP ATCo communicates with two TWR, two other APP controllers, and three ACC (sector) controllers (assuming that there are three entry/exit points of the TMA into the upper area).
- ACC ATCo communicates with three other ACC controllers and one APP ATCo.

Only the TWR and GND ATCo are assumed to work together in the same room. If it is assumed that the amount of communication/co-ordination is equal for all combinations of ATCos, the probabilities as listed in Table 5 can be derived.

Table 5: Communication/coordination distribution

ATCo	Communication/ co-ordination in the same room (**without** use of telephone/radio)	Communication/ co-ordination **not** in the same room (**with** use of telephone / radio)
GND	1.00	0.00
TWR	0.67	0.33
APP	0.00	1.00
ACC	0.00	1.00

6.1.6 Visibility procedure

Five classes of visibility conditions are distinguished:

1. Normal operations: Runway visual range (RVR) above 1500m and cloud base above 300ft.
2. BZO A: RVR between 550m and 1500m or cloud base between 200ft and 300ft (associated with Cat I approach and landings)
3. BZO B: RVR between 350m and 550m or cloud base between 100ft and 200ft (associated with Cat II approach and landings)
4. BZO C: RVR between 200m and 350m (associated with Cat 3A approach and landings)
5. BZO D: RVR below 200m (associated with Cat 3B approach and landings).

A graphical representation of these classes is given in Figure 8.

<table>
<tr><td>300 ft < cloud base</td><td colspan="3">N/A</td><td rowspan="3">2</td><td>1</td></tr>
<tr><td>200 < cloud base < 300ft</td><td colspan="3">2</td><td>2</td></tr>
<tr><td>100 < cloud base < 200ft</td><td colspan="2">3</td><td rowspan="2">3</td><td rowspan="2">N/A</td></tr>
<tr><td>cloud base < 100ft</td><td>5</td><td>4</td><td>N/A</td></tr>
<tr><td></td><td>RVR < 200m</td><td>200 < RVR < 350m</td><td>350 < RVR < 550m</td><td>550 < RVR < 1500m</td><td>RVR > 1500m</td></tr>
</table>

Figure 8: Graphical representation of visibility classes

Because information on the combined occurrence of cloud base and visibility conditions could not be retrieved, the following simplifying assumption was made:

RVR < 200 m visibility class 5

200 < RVR < 350 m visibility class 4
350 < RVR < 550 m visibility class 3
550 < RVR < 1500 m visibility class 2
RVR > 1500 m visibility class 1.

Visibility information was derived from 27 million observations across Europe. A cumulative probability distribution is shown in Figure 9.

Based on these data, the following probabilities are estimated.

Table 6: Probability of occurrence per visibility class

Class	1	2	3	4	5
Probability	0.9587	0.0224	0.0048	0.0033	0.0107

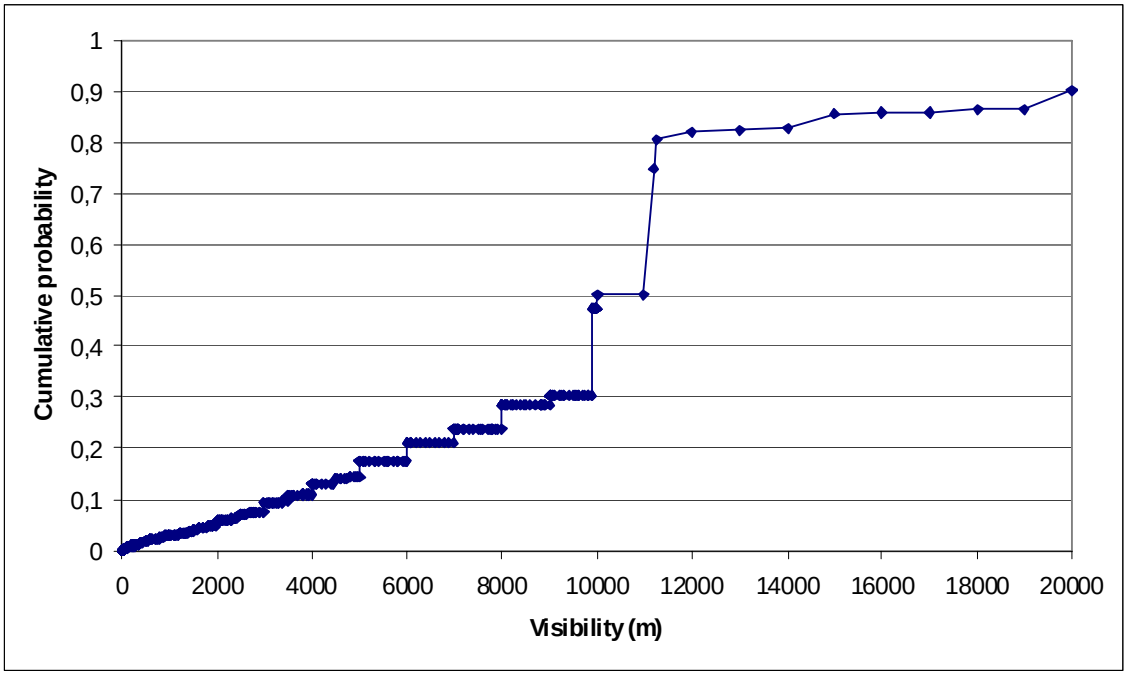


Figure 9: Cumulative probability of visibility

6.2 Expert elicitation

Dependencies between variables were obtained from expert elicitation. The elicitation protocol that was developed for this purpose is represented in Appendix B. Experts were selected on the basis of experience as an air traffic controller. A total of 6 experts were consulted, see Appendix A for a list of names. All experts are air traffic controllers at the Maastricht Upper Area Control Centre (MUAC). The MUAC is responsible for air traffic control in the upper airspace of the Benelux and the north-west of Germany. The dependencies are only considered valid for radar controllers.

In addition to asking questions on the variables of interest, the protocol also included questions on variables for which the correct answers were known to the researchers but not to the experts. Those questions are calibration questions and are important for assessing the performance of the combined experts' assessment. They also form an important part of the feedback to the experts, helping them to gauge their subjective sense of uncertainty against quantitative measures of performance [Cooke and Goossens 2000]. The calibration questions used in this study are presented in Appendix C. Results of the elicitation process are presented in Appendix D. Appendix E describes in detail how the results of the experts were combined.

7 Quantified model

Figure 7 graphically presents the combined model including the marginal distributions for the variables. Dependencies are indicated by conditional and unconditional rank correlations. In Figure 7 these dependencies are represented by their formal notation. For instance $r_{6,1}$ represents the dependence between 'ATCo error' and 'traffic', and $r_{6,2|1}$ represents the dependence between 'ATCo error' and 'man machine interface', given 'traffic'.

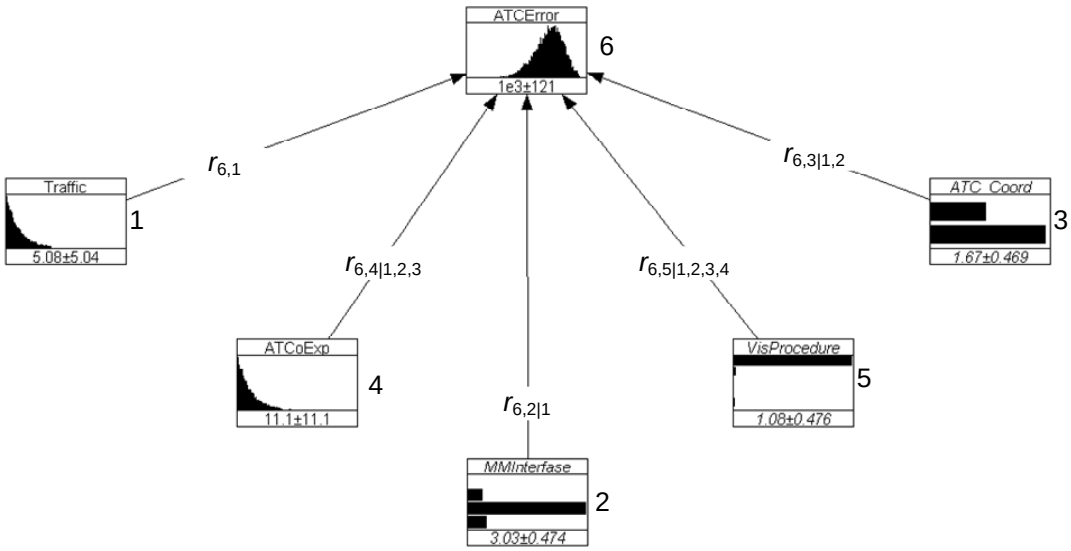


Figure 10: Air traffic controller model

The dependence information resulting from the expert elicitation is presented in Table 8. Individual expert's results were combined using a Global Weight Decision Maker (GWDM), details are provided in Appendix E. All rank correlations were computed from conditional probabilities as in [Morales et al 2007].

Table 8: GWDM dependence information

Probability[5]		(Un)conditional rank correlation	
P_1	0.48	$r_{v_6,v_4\|v_1,v_2,v_3}$	-0.060
P_2	0.44	$r_{v_6,\|v_1}$	-0.179
P_3	0.47	$r_{v_6,v_2\|v_1}$	-0.210
P_4	0.58	$r_{v_6,v_3\|v_1,v_2}$	0.180
P_5	0.52	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.020

8 Conclusions and recommendations

A generic model has been developed with the objective to represent performance of air traffic controllers in a causal model of air transport. Quantification of the model was one of the requirements, therefore the values of all variables are expressed in objectively quantifiable units. The model is built as a Bayesian Belief Network with a combination of continuous and discrete variables. Two versions of the model are proposed; representing a tower controller and a radar controller. The structure of both models is similar, the only difference is the quantification of the dependencies. The radar controller model was fully quantified by a combination of data analysis and expert judgement. A possible future link with managerial influences was anticipated by selecting representatives of different delivery systems[6] as influencing factors of flight crew performance.

The focus of this study was development of a practicable and working model, ready for use within the causal risk model. Consequently and in accordance with the requirements the model is relatively simple. Some of the influences might require further refinement at a later stage, for instance by adding more performance shaping factors.

Quantification of the marginal distributions of the parent nodes and the dependencies is based on readily available data and the opinion of 5 experts. It is recommended to verify the numbers by comparison with alternative sources of data and consulting additional experts. In particular, it is recommended to conduct a separate expert elicitation session to quantify the tower controller model.

[5] P_1 = P(ATCo error ≥ median | Experience ≥ median), P_2 = P(ATCo error ≥ median | Traffic ≥ median), P_3 = P(ATCo error ≥ median | MMI ≥ 4), P_4 = P(ATCo error ≥ median | Communication & Coordination = 2), P_5 = P(ATCo error ≥ median | Visibility Procedure ≥2)

[6] In the causal risk model, safety management is described as providing the resources and criteria for the frontline workforce to operate safely

References

[1] Abbink, F.J. 1996. Integrated free-flight and 4-D gate-to-gate air traffic management, possibilities, promises and problems, NLR TP 96239U, NLR Amsterdam.

[2] Ale, B.J.M., Bellamy, L.J., Cooke, R.M., Goossens, L.H.J., Hale, A.R., Kurowicka, D., Roelen, A.L.C., Smith, E. 2005. Development of a causal model for air transport safety, Proceedings of IMECE2005, 2005 ASME International Mechanical Engineering Congress and Exposition, November 5-11, 2005, Orlando, Florida USA, IMEC2005-79374.

[3] Ale, B.J.M. 2007. Causal Model for Air Transport Safety, Seventh Interim Report. Delft University of Technology, Delft, the Netherlands.

[4] Ale, B.J.M., Bellamy, L.J., Boom, R. van der, Cooper, J., Cooke, R.M., Goossens, L.H.J., Hale, A.R., Kurowicka, D., Morales, O., Roelen, A.L.C., Spouge, J. 2007. Further development of a causal model for air transport safety (CATS): building the mathematical heart. In Aven & Vinnen (eds), Risk, Reliability and Societal Safety, Taylor & Francis Group, London, UK.

[5] ATIC. 2006. Air Traffic Incident Commission, samenvattend jaaroverzicht 2005, ATIC, Hoofddorp.

[6] Bedford, T.J., Cooke, R.M. 2002. Vines - a new graphical model for dependent Random Variables. Annals of Statistics, Vol. 30, No. 4, p. 1031-1068.

[7] Bellamy, L.J., Papazoglou, I.A., Hale, A.R., Aneziris, O.N., Ale, B.J.M., Morris, M.I., Oh, J.I.H. 1999. Development of an integrated technical and management risk control and monitoring methodology for managing and quantifying on-site and off-site risks, I-Risk main report, May 1999, EU Contract number ENVA CT96-0243..

[8] CAA-NL. 2007. Civil aviation safety data 1991-2005, Civil Aviation Authority the Netherlands, The Hague.

[9] Cooke, R.M. 1991. Experts in uncertainty: opinion and subjective probability in science, Environmental Ethics and Science Policy Series, Oxford University Press.

[10] Cooke, R.M., L.J.H. Goossens, L.J.H. 2000. Procedures guide for structured expert judgement, EURATOM document EUR 18820EN, European Commission, Directorate General for Research.

[11] Es, G.W.H. van, Roelen, A.L.C., Kruijsen, E.A.C., Giesberts, M.K.H. 2001. Safety aspects of aircraft performance on wet and contaminated runways, NLR-TP-2001-216, NLR Amsterdam.

[12] Es, G.W.H. van. 2004. Air-ground communication safety study, an analysis of pilot-controller occurrences. Eurocontrol EATMP study.

[13] Eurocontrol. 2004. Age, experience and automation in European Air Traffic Control, - survey in the ECAC area, HRS/HRP-005-REP-05, Eurocontrol, Brussels, Belgium.

[14] Eurocontrol. 2005. The surveillance strategy for ECAC, edition 2.0, Eurocontrol, Brussels, Belgium.

[15] Eurocontrol. 2006. European Convergence and Implementation Plan – status report for 2005, Eurocontrol, Brussels, Belgium.

[16] Hale A.R, Goossens, L.H.J., Bellamy, L.J. 1999. Modelling errors of commission: applicability of ATHEANA in the chemical industry. In: L.H.J. Goossens (eds.); Proceedings Rotterdam '99. Risk Analysis: Facing the New Millennium (Rotterdam, 10/10/99), Delft University Press, Delft, p. 243-246. ISBN: 90-407-1954-3.

[17] Hanea, A.M., Kurowicka, D., Cooke, R.M. 2006. Hybrid method for quantifying and analysing Bayesian belief nets, Quality and Reliability Engineering Journal, Vol. 22, Issue 6, p 709-729.

[18] Hollnagel, E. 1998. Cognitive reliability and error analysis method, Elsevier Science, Oxford.

[19] Kirwan, B. 1994. A practical guide to human reliability assessment, Taylor and Francis, London, UK.

[20] Kirwan, B. 2001. The role of the controller in the accelerating industry of air traffic management, Safety Science, Vol. 37, p. 151-185.

[21] LVNL. 2007. Jaarverslag 2006. Luchtverkeersleiding Nederland, Schiphol-Oost.

[22] Morales, O., Kurowicke, D., Roelen, A. 2007. Eliciting Conditional and Unconditional Rank Correlations from Conditional Probabilities, to be published in Reliability Engineering and System Safety, special issue on expert judgement.

[23] Roelen, A.L.C., Wever, R. 2005. Accident scenarios for an integrated aviation safety model, NLR-CR-2005-560, NLR Amsterdam.

[24] Roelen, A.L.C., Doorn, B.A. van, Smeltink, J.M., Verbeek, M.J., Wever, R. 2006. Quantification of Event Sequence Diagrams for a causal risk model of commercial air transport, NLR-CR-2006-520, NLR Amsterdam.

[25] Roelen, A.L.C., Baren, G.B. van, Smeltink, J.W., Lin, P.H., Morales, O. 2007. A generic flight crew performance model for application in a causal model of air transport. NLR-CR-2007-562, NLR Amsterdam.

[26] Rognin, L, Blanquart, J.-P. 2001. Human communication, mutual awareness and system dependability, lessons learnt from air traffic control field studies, Reliability Engineering and System Safety, Vol. 71, p. 327-336.

[27] Shorrock, S.T., Kirwan, B. 2002, Development and application of a human error identification tool for air traffic control, Applied Ergonomics, Vol. 33, p. 319-336.

[28] Shorrock, S.T. 2005. Errors of memory in air traffic control, Safety Science, Vol 43, p. 571-588.

[29] Shorrock, S.T. 2006 Errors of perception in air traffic control. Safety Science, doi:10.1016/j.ssci.2006.08.018.

[30] Subotic, B., Ochieng, W.Y. Straeter, O. 2007. Recovery from equipment failures in ATC: Determination of contextual factors, Reliability Engineering and System Safety, Vol. 92, p. 858-870.

[31] Swain, A.D., Guttmann, H.E. 1983. Handbook of human reliability analysis with emphasis on nuclear power plant applications, Final report, NUREG/CR-1278-F, SAND80-0200, Sandia National Laboratories, Albuquerque, NM, USA.

[32] U.S. Nuclear Regulatory Commission. 2000. Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA). NUREG-1624, Rev. 1, U.S. Nuclear Regulatory Commission, Washington, DC.

[33] Wever, R. Es, G.W.H. van, Verbeek, M.J. 2005. Air-ground communication safety study: causes and recommendations, NLR-CR-2005-395, NLR Amsterdam.

Appendix A List of experts

Chris Adams
Evelyn Eyres
Paul Fleming
Philip Marien
Steve Mention
Andrew Sovyioltzis

Appendix B Elicitation protocol

B.1 Introduction

Thank you for participating in this expert judgment exercise of the probabilistic characterization of the performance of Air Traffic Controllers. Within the CATS project, in which NLR and TU Delft participate, a model has been developed to represent the causal factors that are supposed to influence the probability of making errors by ATCos.

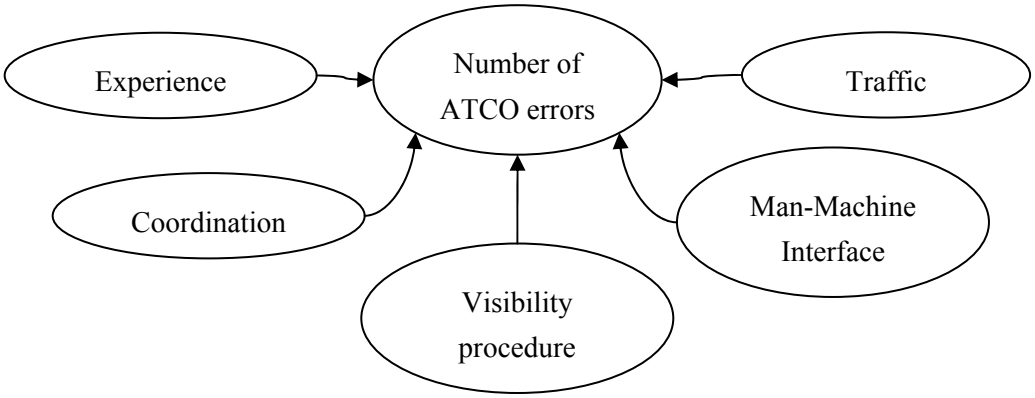
B.2 Objective

The objective of this exercise is to gather information on the variables that influence the number of errors that an ATCo makes in a shift. The potential variables are traffic, visibility, man machine interface, coordination, etc. This information will be used to quantify the ATCo performance model.

In the current version of the model we assume that the influence of a certain variable is independent of the values of other variables.

B.3 The ATCo performance model

The picture below displays the structure of the ATCo performance model. It obviously is a simplified representation of how the number of ATCo errors depends on certain factors. However, these five influential factors (Experience, Traffic, Coordination, Visibility procedure, and Man Machine Interface) are considered to be most important.



The model structure itself is not subject of discussion in the current exercise, and should be considered as it is.

The definition of the variables is given in the table below.

Variable	Definition	Unit
ATCo error probability	Number of errors (not immediately recovered) that an ATCo makes during a shift (the duration of a shift may be different per ATCo role). Example errors are: - ATCo provides pilot with wrong information/instruction - ATCo fails to detect readback error - ATCo provides pilot with an incorrect clearance - ATCo uses incorrect phraseology - ATCo mixes up two aircraft	number per shift
Experience	Number of years working as an ATCo at current position	number of years
Traffic	Number of aircraft simultaneously under control by ATCo	number of aircraft
Man Machine Interface (MMI)	Status of the interface that is used by the ATCo. Four states are distinguished, based on the availability of supporting systems: 1. radio only 2. radio and primary radar 3. radio and primary and secondary radar 4. radio and primary and secondary radar and additional tools (e.g. RIASS, STCA)	1 – 4
ATC communication/ co-ordination	Location of other ATCos with whom communication/coordination is required. 1. in the same room (without use of telephone/radio) 2. not in the same room (with use of telephone/radio)	1 – 2
Visibility procedure	Visibility procedure in use, distinguishing: 1. normal operations: RVR above 1500m and cloud base above 300ft. 2. BZO A: RVR between 550m and 1500m or cloud base between 200ft and 300ft (associated with Cat I) 3. BZO B: RVR between 350m and 550m or cloud base below 200ft (associated with Cat II) 4. BZO C: RVR between 200m and 350m (associated with Cat 3A) 5. BZO D: RVR below 200m (associated with Cat 3B) BZO A – D are in place at Schiphol airport. BZO stands for 'Bijzondere Zicht Operatie' (Limited Visibility Operations). RVR stands for Runway Visual Range.	1 – 5

B.4 Elicitation Method

The method employed is "structured expert judgment", in which the judgment of operational experts on the values of parameters is gathered in a structured way.

Two kinds of information are gathered:

- information on the probability distributions of variables
- information on the influence of variables on the number of errors an ATCo makes during a shift.

Information on the probability distributions of variables

The information on the probability distributions of variables will be elicited via three numbers:

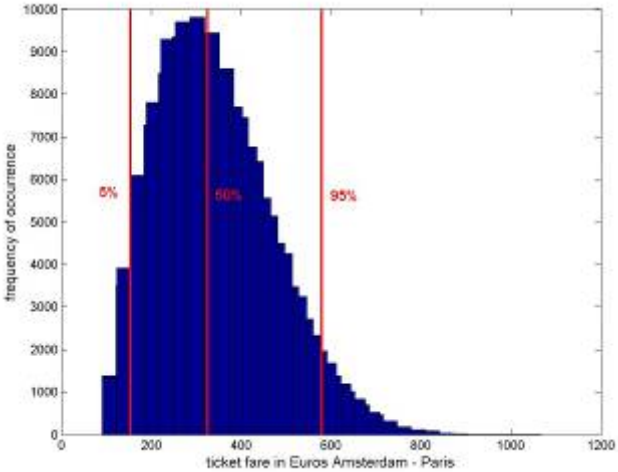
- The median value (50% percentile) of the distribution, i.e. if we would have 101 samples of the variable value, 50 of them should be below and 50 should be higher than the median value.
- The 5% percentile value, which can be interpreted as: it would surprise you if the variable has a value *lower* than this value.
- The 95% percentile value, which can be interpreted as: it would surprise you if the variable has a value *higher* than this value.

As an example, consider the following question:

Consider 100,000 people travelling from Amsterdam to Paris by air. Please give your best estimate of the 5%, 50%, and 95% percentiles of the distribution of the ticket fare (including taxes, considering all classes and airlines) that these 100,000 people have paid.

Your answer could e.g. have been that you estimate that the median value will be 324 Euros, and it would surprise you if there are ticket prices lower than 100 Euros or higher than 500 Euros.

To illustrate the meaning of these percentile values, assume that we have gathered a set of data with 100,000 values of ticket fares. We can make a histogram of the ticket fares as shown in the picture below (this picture is based on artificial data, no real data). The 5%, 50% (median), and 95% percentile values of this data set are indicated by the red lines.



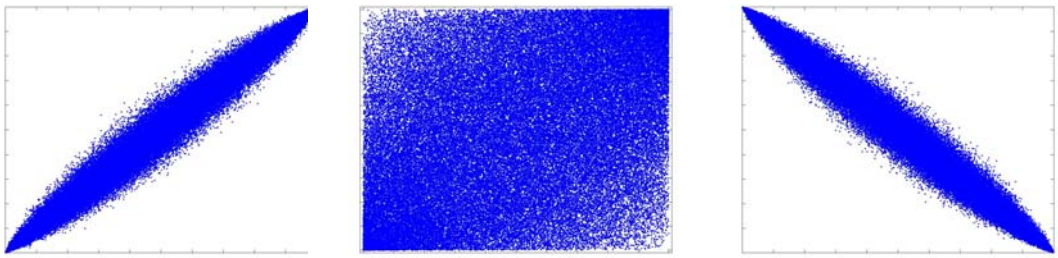
Information on the influence of variables on the number of errors an ATCo makes during a shift

The purpose of the elicitation of dependencies is to obtain 'rank correlation coefficients', which we will refer to as 'correlation' in the following.

Correlation tells something about the relationship between two variables. It can have a value between -1 and +1.

- If *high* values of a variable imply *high* values of another variable, there is a (strong) positive correlation, and the correlation will be close to +1 (the stronger the correlation, the closer it is to +1).
- If *high* values of a variable imply *low* values of another variable, there is a (strong) negative correlation, and the correlation will be close to -1.
- If there appears to be no such relation – the values of a variable do not imply anything on the values of another variable – the variables are called 'independent' and the correlation equals 0.

(Strong) positive *Weak positive* *(Strong) negative*



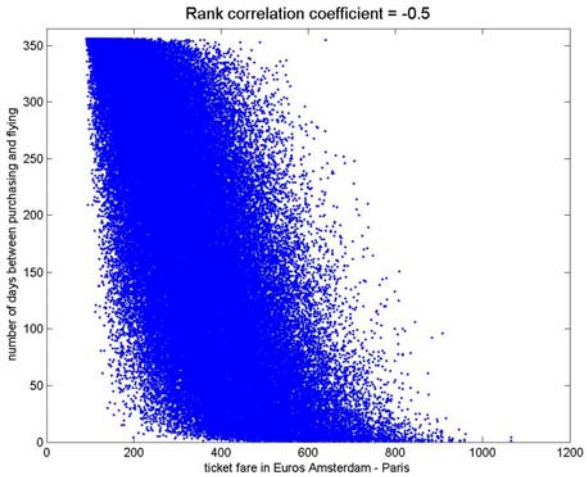
Going back to the example of tickets for Amsterdam – Paris, let's consider three variables that might have an influence on the ticket fare:

- travel class (economy, business, first class)
- time between purchasing and flying (between 0 and 365 days)
- day of the week at which the flight takes place (1=Monday ... 7=Sunday)

Now, what would you expect about the correlation of each of these variables and the ticket fare? Presumably this:

- a higher travel class will result in a higher ticket fare (positive correlation);
- a higher number of days between purchasing and flying will result in a lower ticket fare (negative correlation);
- the day of the week will hardly have an influence on the ticket fare (correlation close to zero).

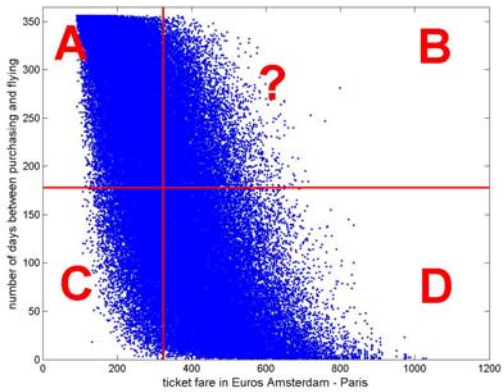
If in addition to the ticket fare we would have gathered the information on the three variables for all 100,000 tickets, we could have made plots like the one below, showing the relations between ticket fare (on the horizontal axis) and the number of days between purchasing and flying (based on artificial data).



As it may sometimes be difficult to directly estimate the correlation, we will use another way of asking, like:

Suppose that out of the 100,000 tickets, we would have selected 50,000 for which the number of days between purchasing and flying is at least equal to its median value (178 days); what portion of these tickets do you expect to have a ticket fare above its median value (above 324 Euros)?

Referring to the picture below, this question comes down to
"*What portion of the blue dots above the red horizontal line lies at the right hand side of the vertical red line (in area B)?"*



- A portion of 50% means that the mass (number of dots) in both areas A and B is equal: there is hardly a relation between the two variables (and the correlation will be close to 0);
- A portion between 0 and 50% implies that there is more mass in A than in B: a *higher* number of days between purchasing and flying will more likely result in a *lower* ticket fare: thus a negative relation (and a correlation between -1 and 0);
- A portion between 50% and 100% implies that there is less mass in A than in B: a *higher* number of days between purchasing and flying will more likely result in a *higher* ticket fare: thus a positive relation (and a correlation between 0 and 1);

We will use this way of asking only for one of the variables. We will pick the variable that you think has the strongest influence on the number of errors that an ATCo makes per shift. For the other variables we will ask for the "relative amount of influence".
Going back to the example of tickets for Amsterdam – Paris once again, this would mean that we first need to order the variables based on the absolute value of the correlation, i.e., make a ranking based on the influence that we think the variables will have, disregarding the 'sign' (positive or negative) of the correlation.
From our example data set we found that the correlation of the ticket fare with travel class was 0.84, with number of days between purchasing and flying was -0.50, and with day of the week was 0.19. So the order would be (from strongest to weakest):

1. travel class (absolute correlation = 0.84)
2. number of days between purchasing and flying (absolute correlation = 0.50)
3. day of the week (absolute correlation = 0.19)

Roughly speaking, one could say that the "relative amount of influence" that the second variable has compared to the first one is 60% (~0.50/0.84) and of the third variable compared to the first one is 20% (~0.19/0.84).
Note that the 'total amount of influence' is limited. Starting from a certain correlation for the variable with the 'strongest influence', it may not be possible for all other variables to have the same amount of influence and at some point, there may be no more influence left over for a certain variable.
You may think of the 'total amount of influence' as a pie that needs to be distributed over a number of persons. If the strongest one takes a large part of it, the rest of the people have to distribute the remaining small chunk amongst them.
During the exercise, we will check if your estimates for the "relative amount of influence" of the variables are still valid in terms of not exceeding the "total amount of influence".

B.5 Confidentiality/feedback

After the individual elicitation, we will report the results including the underlying argumentation and any other discussions and these will be send to you for review. The names and qualifications of the experts will be published in the final report, as will the individual expert assessments and all information relevant, but the link between individual expert assessments and their identity will be removed in the published reports.

Part II: Elicitation of probability distributions

All questions in this section refer to a population of Air Traffic Controllers operating in the Western world (Europe, North America, Australia) and Western-built large aircraft (>5,700 kg Maximum Take-off Weight) currently flying in commercial operations *worldwide*.

B.6 Marginal distribution of number of ATCo errors.

When we speak about ATCo errors, we think of errors that are not immediately recovered, e.g.,

- ATCo provides pilot with wrong information/instruction
- ATCo fails to detect readback error
- ATCo provides pilot with an incorrect clearance
- ATCo uses incorrect phraseology
- ATCo mixes up two aircraft

Q0	***Shift duration***
The number of errors (not immediately recovered) that an ATCo makes are expressed per shift. The duration of a shift may vary from place to place, from time to time, or from position to position. What is your estimate of the average duration in hours of a shift in your current position?	

Q1	***Number of ATCo errors***	
Consider 100,000 ATCos. What are your estimates for the 5th, 50th, and 95th percentiles of the distribution of number of errors that each ATCo makes during a shift?		
_____	_____	_____
5%	50%	95%

B.7 Dependence information.

In this section, we are interested in the relationship between the variable **Number of ATCo errors** and the variables 1-5 in Table 1 (repeated below).

Please assign number 1 to the variable that you consider to have the strongest influence (highest absolute rank correlation coefficient) on the number of ATCo errors, 2 to the next most important and so on until 5 to the least influential variable.

Variable	Definition	Rank
Experience	Number of years working as an ATCo at current position	
Traffic	Number of aircraft simultaneously under control by ATCo	
Man Machine Interface (MMI)	Status of the interface that is used by the ATCo. Four states are distinguished, based on the availability of supporting systems: 1. radio only 2. radio and primary radar 3. radio and primary and secondary radar 4. radio and primary and secondary radar and additional tools (e.g. RIASS, STCA)	
ATC communication/ co-ordination	Location of other ATCos with whom communication/coordination is required. 1. in the same room (without use of telephone/radio) 2. not in the same room (with use of telephone/radio)	
Visibility procedure	Visibility procedure in use, distinguishing: 1. normal operations: RVR above 1500m and cloud base above 300ft. 2. BZO A: RVR between 550m and 1500m or cloud base between 200ft and 300ft (associated with Cat I) 3. BZO B: RVR between 350m and 550m or cloud base below 200ft (associated with Cat II) 4. BZO C: RVR between 200m and 350m (associated with Cat 3A) 5. BZO D: RVR below 200m (associated with Cat 3B) BZO A – D are in place at Schiphol airport. BZO stands for 'Bijzondere Zicht Operatie' (Limited Visibility Operations). RVR stands for Runway Visual Range.	

For the variable that you ranked highest, we have the following question:

Q2	
Suppose that 20,000 **Air Traffic Controllers** are randomly chosen from our total population. Out of those, 10,000 are selected for which the chosen variable (namely: .....) has values above its median value (or above a certain percentile). What portion of these 10,000 **ATCos** will commit more than the median number of errors per shift, (where the median number of ATCo errors was specified in question ***Q1***)?	

For the remaining variables, please indicate their influence as a portion of the influence of the variable that you ranked highest:

Node	**Rank**	**Influence as a portion of the influence of the highest ranked variable (0 – 100%)**	**Direction of the correlation (positive/negative)**
Experience			
Traffic			
Man Machine Interface (MMI)			
ATC communication/ co-ordination			
Visibility procedure			

Part III: Calibration Variables

All questions in this section refer to a population of Air Traffic Controllers operating in the Western world (Europe, North America, Australia) and Western-built large aircraft (>5,700 kg Maximum Take-off Weight) currently flying in commercial operations *worldwide*.

To capture your uncertainty, in all questions from this section, we will ask you to provide the 5%, 50% and 95% percentiles of your uncertainty distribution, which can be interpreted as that we ask for your best estimate (50%), the value which would surprise you if the real value would be lower (5%), and the value which would surprise you if the real value would be higher (95%).

CQ1		
What was the number of airprox reports in The Netherlands in 2005?		
5%	50%	95%

CQ2		
What is the missed approach rate at Schiphol (based on RWY 06 data)?		
5%	50%	95%

CQ3		
What was the number of infringements (unauthorized entries) of controlled airspace in Amsterdam FIR in 2006?		
5%	50%	95%

CQ4		
What was the size of the world fleet of large (100 passenger seats or more) commercial jet in 2003		
5%	50%	95%

CQ5	**Not used**

CQ6		
What percentage of communication problems between ATCo and pilot is related to a lacking pilot readback (between March 2004 and April 2005 in European airspace)?		
5%	50%	95%

CQ7	**Not used**

CQ8		
Out of the total number of onboard fatalities, what percentage is due to a runway overrun accident (worldwide data 1991-2005)?		
5%	50%	95%

CQ9		
What percentage of the annual number of take-offs and landings in The Netherlands is performed on a wet or contaminated runway (based on data from 1976-1995)?		
5%	50%	95%

CQ10		
What part of flights at Schiphol is operated with first generation aircraft (B707, etc.) in 2005?		
5%	50%	95%

CQ11		
What is the probability for a runway overrun during take-off per flight at Schiphol?		
5%	50%	95%

CQ12		
What is the probability for a runway incursion with a significant potential for a collision (Class A or B incursion) per flight (in the ECAC area in 2002)?		
5%	50%	95%

Appendix C Answers to calibration questions.

CQ1: What was the number of airprox reports in the Netherlands in 2007?
Answer: 25
Source: ATIC report 2005 [ATIC 2006].

CQ2: What is the missed approach rate at Schiphol (based on RWY 06 data)?
Answer: 1.10×10^{-3}
Source: Air Traffic Control the Netherlands.

CQ3: What was the number of infringements (unauthorized entries) of controlled airspace in the Amsterdam FIR in 2006?
Answer: 214
Source: Air Traffic Control the Netherlands, annual report 2006 [LVNL 2007].

CQ4: What was the size of the fleet of large (100 seats or more) commercial jet aircraft in 2003?
Answer: 10838
Source: Airbus market forecast 2004-2023, page 43.
http://www.airbus.com/en/myairbus/global_market_forcast.html

CQ6: What percentage of communication problems between ATCo and pilot is related to a lacking pilot readback (between March 2004 and April 2005 in European airspace)?
Answer: 1
Source: NLR report 'Air ground communication safety study' [Wever et al 2005].

CQ8: Out of the total number of on-board fatalities, what percentage is due to a runway overrun accident (wordwide data 1991-2005)?
Answer: 6
Source: Civil aviation safety data 1991-2005 [CAA-NL 2007].

CQ9: What percentage of the annual number of take-offs and landings in the Netherlands is performed on a wet or contaminated runway (based on data from 1976-1995)?
Answer: 20
Source: NLR report 'Safety aspects of aircraft performance on wet and contaminated runways' [Van Es et al 2001].

CQ10: What part of flights at Schiphol is operated with first generation aircraft (Boeing 707, etc) in 2005?
Answer: 0.23
Source: Civil aviation safety data 1991-2005 [CAA-NL 2007].

CQ11: What is the probability for a runway overrun during take-off per flight at Schiphol?
Answer: 6.50×10^{-8}
Source: NLR report 'Quantification of Event Sequence Diagrams for a causal risk model of commercial air transport' [Roelen et al 2006].

CQ12: What is the probability for a runway incursion with a significant potential for collision (Class A or B incursion) per flight (in the ECAC area in 2002)?
Answer: 1.60×10^{-5}
Source: NLR report 'Quantification of Event Sequence Diagrams for a causal risk model of commercial air transport' [Roelen et al 2006].

Appendix D Expert's results for probability distributions

Table D.1: Experts' results for ATCo error distribution

	Variable		A	B	C	D	E
Q0	Shift duration		2.5	1.5	1.75	2	1.5
Q1	Number of ATCo error	5%	1	1	2	1	0.1
		50%	7	3	3	2	0.6
		95%	20	5	5	4	0.95

Table D.2: Experts' results for calibration questions

	Calibration variable		Data	A	B	C	D	E
CQ1	Airprox reports	5%		1	15	10	8	10
		50%	25	6	25	15	13	40
		95%		2	30	30	15	60
CQ2	Missed approach rate	5%		0	$1.0x10^{-2}$	$1.0x10^{-2}$	$3.0x10^{-2}$	$2.0x10^{-3}$
		50%	$1.1x10^{-3}$	$1.0x10^{-4}$	$3.0x10^{-2}$	$2.0x10^{-2}$	$1.0x10^{-1}$	$2.0x10^{-2}$
		95%		$2.0x10^{-3}$	$1.0x10^{-1}$	$5.0x10^{-2}$	$2.0x10^{-1}$	$5.0x10^{-2}$
CQ3	Infringements	5%		5	5	4	250	5
		50%	214	150	12	10	360	80
		95%		900	32	10	500	150
CQ4	World fleet size	5%		1,000	5,000	3,000	20,000	30,000
		50%	10,838	30,000	10,000	6,000	20,000	60,000
		95%		65,000	50,000	7,000	30,000	200,000
CQ6	Pilot readback error	5%		5	30	20	40	5
		50%	1	65	70	30	60	13
		95%		95	90	50	70	30
CQ8	Runway overrun fatalities	5%		0.1	1	3	1	2
		50%	6	6	5	10	2	15
		95%		50	40	15	10	50
CQ9	Wet runway movements	5%		2	10	10	35	30
		50%	20	40	40	55	55	45
		95%		99	60	70	75	60
CQ10	Percentage 1st generation	5%		0.1	5	2	10	3
		50%	0.23	5	10	5	15	8
		95%		40	25	8	25	12
CQ11	Overrun probability	5%		$1.0x10^{-6}$	0	$1.0x10^{-6}$	$3.0x10^{-6}$	$1.0x10^{-4}$
		50%	$6.5x10^{-8}$	$1.0x10^{-5}$	$1.0x10^{-2}$	$2.5x10^{-6}$	$1.0x10^{-5}$	$1.0x10^{-3}$
		95%		$1.0x10^{-4}$	$5.0x10^{-2}$	$1.0x10^{-5}$	$2.0x10^{-5}$	$5.0x10^{-2}$
CQ12	Runway incursion probability	5%		$1.0x10^{-6}$	0	$2.5x10^{-6}$	$1.0x10^{-2}$	$1.0x10^{-5}$
		50%	$1.6x10^{-5}$	$2.0x10^{-6}$	$1.0x10^{-2}$	$1.0x10^{-5}$	$5.0x10^{-2}$	$1.0x10^{-4}$
		95%		$4.0x10^{-6}$	$5.0x10^{-2}$	$3.0x10^{-5}$	$1.0x10^{-1}$	$5.0x10^{-3}$

Appendix E Description of the expert elicitation results

E.1 Quantification of the ATCo performance model

This appendix reports on the final quantification of the ATCo performance model, see Figure E.1. A summary of the variables entered in the model is presented in Table E.1. The (un)conditional rank correlations shown in Figure E.1 have been elicited through one conditional probability of exceedance [Morales et al 2007] and ratios of unconditional rank correlations. These have been translated to conditional probabilities of exceedance later and combined through the classical method for expert judgment [Cooke 1991]. In total, six experts have been elicited but only five used for the analyses. Expert number six presented estimates that were inconsistent and his/her estimates could not be used. Results of the expert judgment exercise will be discussed in the next sections.

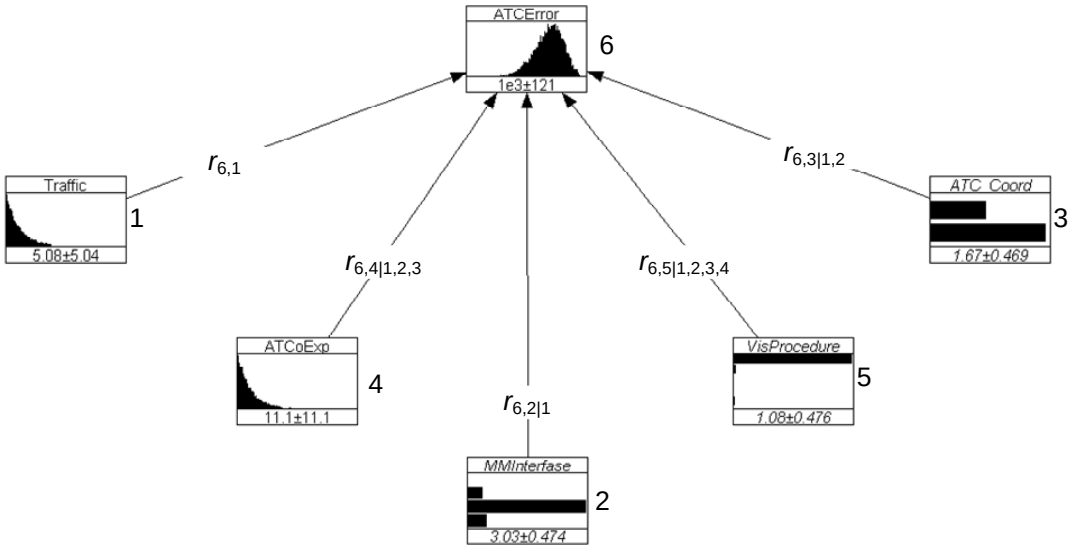


Figure E.1: ATC performance model

Table E.1: ATCo model nodes and their definitions

Node	Definition	Unit	Source for marginal distribution
Experience	Number of years working as an ATCo at current position	Number of years	Eurocontrol report on Age, Experience and Automation in European ATC – Survey in the ECAC area [Eurocontrol 2004]
Traffic	Number of aircraft simultaneously under control by ATCo	Number of aircraft	CFMU data
Man Machine Interface (MMI)	Classification of the interface that is used by the ATCo. Four classes are distinguished, based on the availability of supporting systems: 5. radio only 6. radio and primary radar	1 – 4	General information on European airports [Eurocontrol 2005] and [Eurocontrol 2006].

Node	Definition	Unit	Source for marginal distribution
	7. radio and primary and secondary radar 8. radio and primary and secondary radar and additional tools (e.g. RIASS, STCA)		
ATC co-ordination	Whether or not the communication/ co-ordination between ATCos takes place in the same room. In the latter case it is assumed that telephone or radio is used.	yes/no	Actual situation at Schiphol.
Visibility procedure	Visibility procedure in use, subdivided into 5 classes: 6. normal operations: RVR above 1500m and cloud base above 300ft. 7. BZO A: RVR between 550m and 1500m or cloud base between 200ft and 300ft (associated with Cat I) 8. BZO B: RVR between 350m and 550m or cloud base below 200ft (associated with Cat II) 9. BZO C: RVR between 200m and 350m (associated with Cat 3A) 10. BZO D: RVR below 200m (associated with Cat 3B) BZO A – D are in place at Schiphol airport. BZO stands for 'Bijzondere Zicht Operatie' (Limited Visibility Operations). RVR stands for Runway Visual Range.	1 – 5	Schiphol operational information in combination with meteorological data

E.2 The classical model for expert judgement

The so-called Classical Model [Cooke 1991] for deriving uncertainty distributions over model parameters from expert judgements will be used. The Classical Model is a performance based linear pooling or weighted averaging model. The weights are derived from experts' calibration and information performance, as measured on calibration or seed variables. These are variables from the experts' field whose values become known to the experts post hoc. Seed variables serve a threefold purpose: (i) to quantify experts' performance as subjective probability assessors, (ii) to enable performance-optimised combinations of expert distributions, and (iii) to evaluate and hopefully validate the combination of expert judgements.

Calibration measures the statistical likelihood that a set of experimental results corresponds, in a statistical sense, with the experts' assessments. In particular, the calibration score is the p-value of a standard Chi-square goodness of fit test. Loosely, the calibration score is the probability that the divergence between the expert's probabilities and the observed values of the seed variables might have arisen by chance. A high score (near one, but bigger than, say, 0.05) means that the expert's probabilities are statistically supported by the set of seed variables. Information represents the degree to which an expert's distribution is concentrated, relative to some user-selected background measure. The overall information score is the mean of information scores for each variable.

"Good expertise" corresponds to good calibration (high statistical likelihood) and high information. Highly informative distributions are desirable only if they are well calibrated. Information should discriminate between more or less equally well calibrated experts. The weights in the classical model are proportional to the product of statistical likelihood and information, and it turns out that this product is indeed dominated by calibration. When a combined expert has been formed, we can also measure the calibration and information of this combined expert. For more detail see [Cooke 1991]. Calculations are performed with the EXCALIBUR software developed at TU Delft.

E.3 Expert's performance

Table E.2 shows the calibration and information scores for the five experts used in the study. The first and second columns give the expert number and id respectively; the third column gives the calibration score. The ratio of highest to lowest score is about 3.26 x 10^7. It will be noted that only expert A had a score corresponding to a p-value above 5%. Scores of Experts D and E are marginal and for expert C is rather low. Calibration scores in the order 0.001 would fail to confer the requisite level of confidence in the results.

The information scores for all items and for calibrations items are shown in columns 3 and 4 respectively. It will be noted that the overall information scores are quite similar, within a factor 3. In this case the expert with the best calibration score (Expert A) also has the lowest information score for the calibration variables which is a recurrent pattern. The last column gives the "un-normalized weight"; this is the product of the calibration score and the information score for calibration variables. If this column were normalized and used to form weighted combinations, experts A, B and C would be influential with (52.07, 46.11 and 1.80 per cent respectively).

Table E.2: Expert's performance

Expert ID	Calibration score	Information score for all variables	Information score for calibration variables	Un-normalised weight
A	0.1012	0.5633	0.5034	0.05095
B	0.04706	1.03	0.9588	0.04512
C	0.00131	1.423	1.349	0.001767
D	2.795×10^{-9}	1.669	1.655	4.625×10^{-9}
E	2.501×10^{-6}	1.017	0.9624	2.407×10^{-6}

E.4 Combination schemes

In this study, experts were asked to give their uncertainty assessments on calibration variables in the form of 5%, 50%, and 95% quantiles. To combine all experts' assessments into one uncertainty assessment we can apply different combination schemes. The combined

distributions are weighted sums of the individual experts' distributions, with non-negative weights adding to one. Different combination schemes are distinguished by the method according to which the weights are assigned to densities. These schemes are designated "Decision Makers". Three kinds of decision makers are described below.

Equal Weight Decision Maker

The equal weight decision maker (EWDM) results by assigning equal weight to each density. In Table E.3 the 6th expert is identified as "EWDM". As it may be seen in Table E.3 the EWDM is better calibrated than each expert individually. However information scores derived from the EWDM are poor. They are the lowest amongst the 6 experts (that is including the EWDM as an expert) in both all variables and calibration questions alone. Weights for each expert and for the decision maker are shown in the last three columns. Since the equal weight combination has been chosen in this case, column 6 displays the same weight for all experts except the decision maker, and finally column 7 shows that when normalizing the numbers in column 6 the decision maker's weight is 1.6 times smaller than the best calibrated expert.

Table E.3: Expert's and EW Decision Maker's performance

Expert ID	Calibration score	Information score for all variables	Information score for calibration variables	Un-normalised weight	Normalized weight without DM	Normalized weight with DM
A	0.1012	0.5633	0.5034	0.05095	0.2	0.3964
B	0.04706	1.03	0.9588	0.04512	0.2	0.351
C	0.00131	1.423	1.349	0.001767	0.2	0.01375
D	2.795×10^{-9}	1.669	1.655	4.625×10^{-9}	0.2	3.598×10^{-8}
E	2.501×10^{-6}	1.017	0.9624	2.407×10^{-6}	0.2	1.872×10^{-5}
EWDM	0.1242	0.2662	0.2472	0.0307		0.2388

Item and Global Weight Decision Maker

This kind of decision maker is from the class of performance based decision makers that are those where the weights are based on the experts' performance on calibration variables. Two performance based decision makers are supported in the software EXCALIBUR. The "Global Weight" decision maker and the "Item Weight" decision maker.

The Global Weight Decision Maker (Table E.4) uses average information over all calibration variables and computes one set of weights for all items. It computes performance based weights which are defined, per expert, by the product of expert's calibration score and his(her) overall information score on calibration variables, and by an optimization procedure[7]. In this case, all

[7] The weights used in the classical model are *proper scores*. A score is called *strictly proper* if an expert can achieve his highest expected score by and only by stating his true opinion. The measures for information and calibration described above must be combined in such a way that the result is (in the long run) a strictly proper scoring rule. This requires that the measurement of

experts with a calibration score less than the significance level found by the optimization procedure are outweighed as reflected by the zeros in columns 6 and 7 in Table E.4.

In Table E.4 one can see that after the optimization procedure is applied, 3 experts have non-zero weight. Comparing Tables E.3 and E.4 one can see that the calibration score of the GWDM is about 3 times higher than the EWDM. The information score is comparable for both decision makers in both all variables and calibration variables alone.

Table E.4: Expert's and GW Decision Maker's performance

Expert ID	Calibration score	Information score for all variables	Information score for calibration variables	Un-normalised weight	Normalized weight without DM	Normalized weight with DM	Information relative to DM for all variables	Information relative to DM for calibration variables
A	0.1012	0.5633	0.5034	0.05095	0.5208	0.2015	0.379	0.4075
B	0.04706	1.03	0.9588	0.04512	0.4612	0.1784	0.578	0.5823
C	0.00131	1.423	1.349	0.001767	0.01806	0.006987	1.135	1.156
D	2.795×10^{-9}	1.669	1.655	4.625×10^{-9}	0	0	1.458	1.55
E	2.501×10^{-6}	1.017	0.9624	2.407×10^{-6}	0	0	1.229	0.9386
GWDM	0.6827	0.3094	0.2271	0.1551		0.6131	0	0

The Item Weight Decision Maker (Table E.5) constructs weights for each item separately, using the experts' information scores for the given item, rather than the average information score. The calibration score of the GWDM is about 2.8 times larger than the IWDM. The IWDM is slightly more informative than the GWDM. However the gain in information is not a sufficient argument to justify a preference of the IWDM over the GWDM.

calibration be combined with "classical significance tests". Briefly, there must be some value $\alpha > 0$, such that if the expert's statistical likelihood drops below α, his/her weight becomes zero. For each value of α it is defined a decision maker dm_α computed as a weighted linear combination of the experts whose calibration score exceeds α. dm_α is scored with respect to calibration and information. The weight which this dm_α would receive if he were added as a "virtual expert" is called the "virtual weight" of dm_α. The value of α for which the virtual weight of dm_α is the greatest is chosen as the cut-off value for determining the unweighted expert.

Table E.5: Expert's and IW Decision Maker's performance

Expert ID	Calibration score	Information score for all variables	Information score for calibration variables	Un-normalised weight	Normalized weight with DM
A	0.1012	0.5633	0.5034	0.05095	0.2688
B	0.04706	1.03	0.9588	0.04512	0.238
C	0.00131	1.423	1.349	0.001767	0.009322
D	2.795×10^{-9}	1.669	1.655	4.625×10^{-9}	0
E	2.501×10^{-6}	1.017	0.9624	2.407×10^{-6}	0
IWDM	0.2441	0.4441	0.3757	0.0917	0.4838

Figure E.2 presents range graphs for the experts' assessments in calibration variables. The GWDM is also shown for comparison. The 5th and 95th percentile of each expert's uncertainty distribution is represented by the error bar. The predominant picture is that the experts' central 90% confidence show considerable overlap. To appreciate Figure E.2, a comparison of each expert to the GWDM is also provided in Table E.4. In discrepancy analysis, the relative information of each expert with respect to the Decision Maker is computed per item. These scores are averaged over all variable (column 8) and calibration variables only (column 9). It can be observed that experts tend to agree with the GWDM. The ratio of column 8 to 9 is close to 1 in each case.

In summary the recommended choice of the decision maker is the Global Weight Decision Maker as it achieves better performance than the Equal Weight and Item Weight combinations. Future analysis will be performed also based on the GWDM.

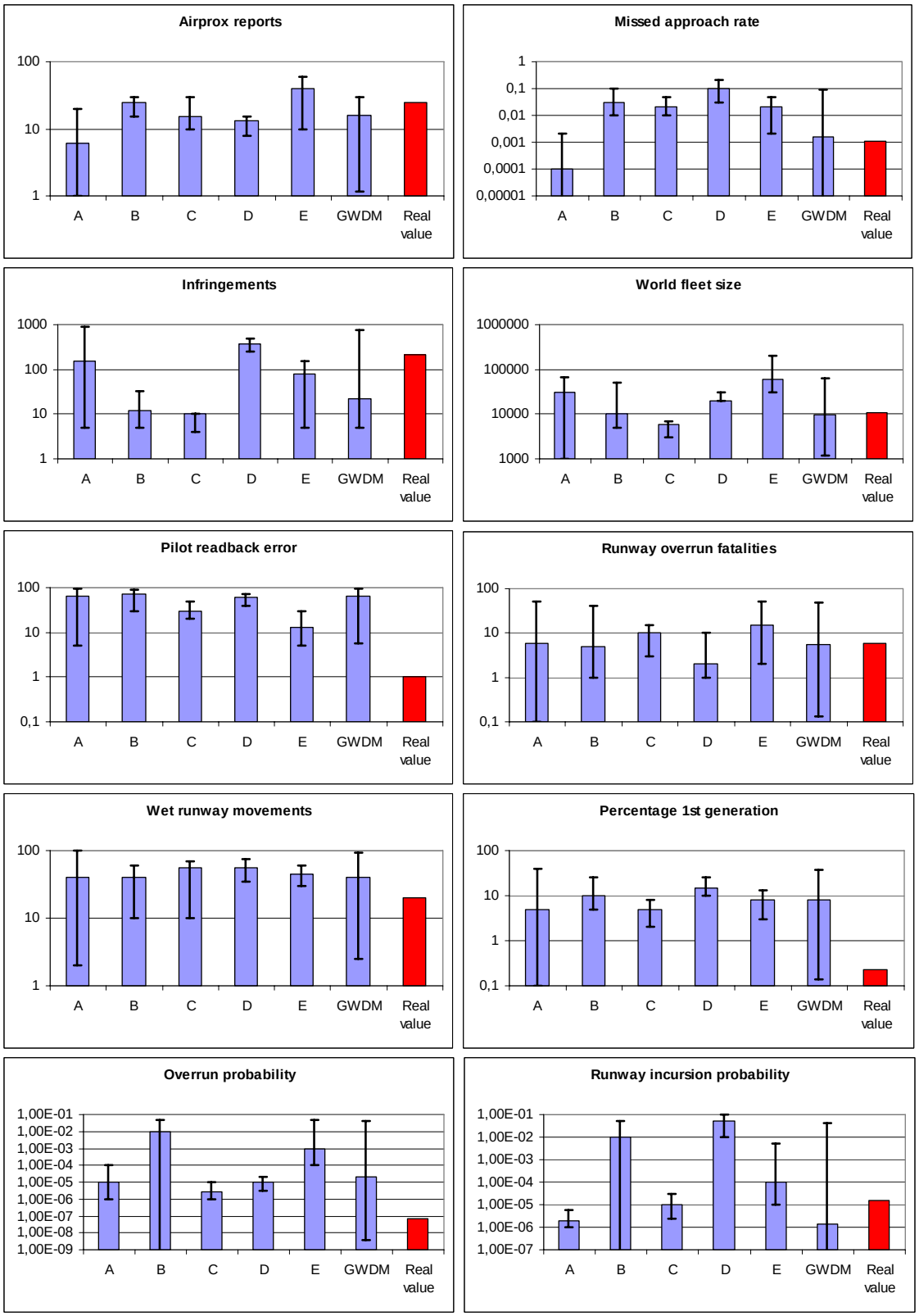


Figure E.2: Range graphs for expert's assessments in calibration variables

E.5 Combining experts' dependence opinions

In the classical method for expert judgement, the marginal distributions of experts $e = 1, \ldots N$ are combined through a weighted average. In equation (E.1) $f_{Y,DM}$ is the density of elicitation variable Y of the DM, w_e the weight for expert e and $f_{Y,e}$ the density of expert e on variable Y. The weights can be chosen as equal weights or derived from performance measures.

$$f_{Y,DM} = \sum_{e=1}^{N} w_e f_{Y,e} \tag{E.1}$$

From previous analysis it was concluded that the GWDM is the combination of choice. The combination of experts' dependence estimates derives from equation (E.1) above with some modifications. A questionnaire (Appendix B) was designed to elicit the dependence information from Figure E.1. This questionnaire included 10 calibration variables. Additionally 6 questions were asked to obtain the rank and conditional rank correlations required by the model in Figure E.1. These 6 questions were essentially the following:

1. A ranking of variables assigning 1 to the one with the largest influence (highest absolute rank correlation coefficient) on ATCo errors until 5 for the one with the smallest absolute rank correlation. For each expert variables were labelled as *v1,… , v5*. Variable *v6* would denote ATCo errors for all experts. Notice that *v1,… , v5* could denote different variables for each expert according to their own beliefs.
2. A probability of exceedance for the 'most influential' variable was asked to each expert in the usual way. For example: suppose variable *v1* was found to lie above its median value[8] what is your probability that also *v6* would be above its median value. This would give us the rank correlation $r_{v6,vj} / r_{v6,v1}$ for $j = 2,\ldots,5$. Observe that the ratios should be decreasing as j increases.

This information together with the marginal distributions for each node is enough to compute the (un)conditional rank correlations required by the model. The results of asking the 6 questions above are presented in Table E.6. All (un)conditional rank correlations were computed from conditional probabilities of exceedance as in [Morales et al 2007] assuming a normal copula to construct the joint distribution.

[8] Or some other percentile if the variable is discrete.

Table E.6: Expert elicitation results for dependence information of the ATCo model in Figure E.1

A						
Variable	Rank	Ratio	Probability[9]		(Un)conditional rank correlation	
Experience	4	0.20	P_1	0.515	$r_{v_6,v_4\|v_1,v_2,v_3}$	0.05
Traffic	2	-0.40	P_2	0.471	$r_{v_6,v_2\|v_1}$	-0.09
MMI	3	-0.20	P_3	0.494	$r_{v_6,v_3\|v_1,v_2}$	-0.05
Comm. & Coord.	1		P_4	0.6	$r_{v_6,\|v_1}$	0.22
Visibility Procedure	5	0.20	P_5	0.539	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.05

B						
Variable	Rank	Ratio	Probability		(Un)conditional rank correlation	
Experience	3	-0.50	P_1	0.443	$r_{v_6,v_3\|v_1,v_2}$	-0.19
Traffic	2	-0.75	P_2	0.414	$r_{v_6,v_2\|v_1}$	-0.27
MMI	1		P_3	0.45	$r_{v_6,\|v_1}$	-0.34
Comm. & Coord.	4	0.40	P_4	0.562	$r_{v_6,v_4\|v_1,v_2,v_3}$	0.16
Visibility Procedure	5	0.00	P_5	0.5	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.00

C						
Variable	Rank	Ratio	Probability		(Un)conditional rank correlation	
Experience	3	-0.65	P_1	0.374	$r_{v_6,v_3\|v_1,v_2}$	-0.54
Traffic	1		P_2	0.3	$r_{v_6,\|v_1}$	-0.57
MMI	2	-0.70	P_3	0.442	$r_{v_6,v_2\|v_1}$	-0.50
Comm. & Coord.	4	-0.60	P_4	0.341	$r_{v_6,v_4\|v_1,v_2,v_3}$	-0.60
Visibility Procedure	5	0.00	P_5	0.5	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.00

D						
Variable	Rank	Ratio	Probability		(Un)conditional rank correlation	
Experience	4	-0.25	P_1	0.442	$r_{v_6,v_4\|v_1,v_2,v_3}$	-0.28
Traffic	3	0.26	P_2	0.56	$r_{v_6,v_3\|v_1,v_2}$	0.28
MMI	1		P_3	0.4	$r_{v_6,\|v_1}$	-0.69
Comm. & Coord.	2	0.40	P_4	0.628	$r_{v_6,v_2\|v_1}$	0.39
Visibility Procedure	5	0.10	P_5	0.562	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.12

E						
Variable	Rank	Ratio	Probability		(Un)conditional rank correlation	
Experience	1		P_1	0.4	$r_{v_6,\|v_1}$	-0.30
Traffic	3	0.20	P_2	0.52	$r_{v_6,v_3\|v_1,v_2}$	0.07
MMI	2	0.70	P_3	0.531	$r_{v_6,v_2\|v_1}$	0.22
Comm. & Coord.	4	0.10	P_4	0.514	$r_{v_6,v_4\|v_1,v_2,v_3}$	0.03
Visibility Procedure	5	0.03	P_5	0.508	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.01

GWDM						
Variable	Rank	Ratio	Probability		(Un)conditional rank correlation	
Experience		-	P_1	0.48	$r_{v_6,v_4\|v_1,v_2,v_3}$	-0.060
Traffic		-	P_2	0.44	$r_{v_6,\|v_1}$	-0.179
MMI		-	P_3	0.47	$r_{v_6,v_2\|v_1}$	-0.210
Comm. & Coord.		-	P_4	0.58	$r_{v_6,v_3\|v_1,v_2}$	0.180
Visibility Procedure		-	P_5	0.52	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.020

[9] $P_1 = P$(ATCo error ≥ median | Experience ≥ median), $P_2 = P$(ATCo error ≥ median | Traffic ≥ median), $P_3 = P$(ATCo error ≥ median | MMI ≥ 4), $P_4 = P$(ATCo error ≥ median | Communication & Coordination = 2), $P_5 = P$(ATCo error ≥ median | Visibility Procedure ≥2)

After the estimates from each expert were obtained, they were combined with the procedure described in section E.8. The general idea consists basically in taking a weighted average of each expert's individual assessments for a given probabilistic statement concerning the joint distribution. The weights were taken from the GWDM in Table E.4.

Table E.7: GWDM dependence information

Probability[10]		(Un)conditional rank correlation	
P_1	0.48	$r_{v_6,v_4\|v_1,v_2,v_3}$	-0.060
P_2	0.44	$r_{v_6,\|v_1}$	-0.179
P_3	0.47	$r_{v_6,v_2\|v_1}$	-0.210
P_4	0.58	$r_{v_6,v_3\|v_1,v_2}$	0.180
P_5	0.52	$r_{v_6,v_5\|v_1,v_2,v_3,v_4}$	0.020

Table E.7 presents the final results of applying the procedure described in section E.8 for combining dependence information to the expert's individual estimates in Table E.6. The determinant of the correlation matrix of the BBN in Figure E.1 with the estimates in Table E.7 is equal to 0.89 showing that the GWDM considers a very small linear relationship of ATCo error with other variables in Table E.1. This may be explained by observing that experts A and B are dominant in the performance based combination. For expert A the highest rank correlation with ATCo error was communication and coordination equal to 0.22 ($r_{6,1}$ = 0.22 in Table E.6). For expert B the highest was -0.34 with man-machine interface ($r_{6,1}$ = -0.34 in table E.6). This explains the high value of the determinant of the correlation matrix of the GWDM. Expert C has the lowest determinant for the correlation matrix amongst all 5 experts but his/her contribution is marginal to the performance based combination. Experts D and E do not contribute to the GWDM.

E.6 Conclusions

This Appendix E shows that the elicitation and optimal combination of expert judgments in the form of rank and conditional rank correlations for real applications is possible. The optimal combination regards variables 1 – 5 as showing strong non-monotonic relationships with ATCo error.

[10] P_1 = P(ATCo error ≥ median | Experience ≥ median), P_2 = P(ATCo error ≥ median | Traffic ≥ median), P_1 = P(ATCo error ≥ median | MMI ≥ 4), P_1 = P(ATCo error ≥ median | Communication & Coordination = 2), P_1 = P(ATCo error ≥ median | Visibility Procedure ≥2)

E.7 Procedure for combining expert's dependence opinions

To combine the dependence information elicited from expert's via conditional probabilities it would be tempting to pool the conditional probabilities linearly to determine the conditional probability of the decision maker. This strategy would work well in general if the medians off all experts were the same which is not typically the case. In order to combine the experts' dependence information a different strategy has to be taken. An example is presented with the BBN in Figure E.3.

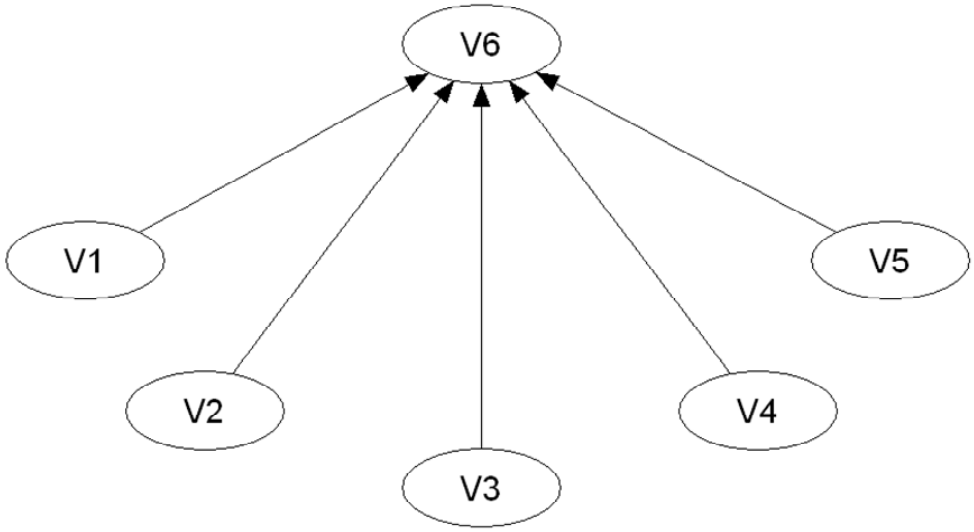


Figure E.3: BBN example for the elicitation procedure

The rank correlations from each expert are elicited as in [Morales et al 2007]. Denote each expert as e_i for i = 1,…,N. The rank correlations will be denoted as $r_{6,1}^{e_i}$ for the rank correlation between V6 and V1 in Figure E.3 for each i. Similarly the conditional rank correlation between V6 and V2 given V1 will be denoted as $r_{6,2|1}^{e_i}$ for expert e_i. All other (un)conditional rank correlations in the BBN will be denoted similarly. The median value of variable V_j for expert e_i is denoted as $v_{j,q_{50}}^{e_i}$. Similarly the kth percentile of variable V_j is denoted as $v_{j,q_k}^{e_i}$. Proceed as follows.

1. Query from each expert e_i some probabilistic statement as in relation (E.2) and translate them to the corresponding rank correlations. $F_{V_j}^{e_i}$ is the cumulative distribution function of V_j for expert e_i. Observe that $v_{j,q_{50}}^{e_i}$ will in general be different for each variable and for each expert when the marginal distribution comes from expert judgment. When the marginal distributions come from data then the median value across experts and the decision maker will be the same.

$$
\begin{aligned}
P_1^{e_i} &= P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_1 \geq v_{1,q_{50}}^{e_i}\right) \\
&= P\left(F_{V_6}^{e_i}(V_6) \geq 0.5 \mid F_{V_1}^{e_i}(V_1) \geq 0.5\right) \\
P_2^{e_i} &= P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_1 \geq v_{1,q_{50}}^{e_i}, V_2 \geq v_{2,q_{50}}^{e_i}\right) \\
&= P\left(F_{V_6}^{e_i}(V_6) \geq 0.5 \mid F_{V_1}^{e_i}(V_1) \geq 0.5, F_{V_2}^{e_i}(V_2) \geq 0.5\right) \\
P_3^{e_i} &= P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_1 \geq v_{1,q_{50}}^{e_i}, V_2 \geq v_{2,q_{50}}^{e_i}, V_3 \geq v_{3,q_{50}}^{e_i}\right) \\
&= P\left(F_{V_6}^{e_i}(V_6) \geq 0.5 \mid F_{V_1}^{e_i}(V_1) \geq 0.5, F_{V_2}^{e_i}(V_2) \geq 0.5, F_{V_3}^{e_i}(V_3) \geq 0.5\right) \\
&\ \ . \\
&\ \ . \\
&\ \ . \\
P_5^{e_i} &= P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_1 \geq v_{1,q_{50}}^{e_i}, V_2 \geq v_{2,q_{50}}^{e_i}, \ldots, V_5 \geq v_{5,q_{50}}^{e_i}\right) \\
&= P\left(F_{V_6}^{e_i}(V_6) \geq 0.5 \mid F_{V_1}^{e_i}(V_1) \geq 0.5, F_{V_2}^{e_i}(V_2) \geq 0.5, \ldots, F_{V_5}^{e_i}(V_5) \geq 0.5\right)
\end{aligned}
\tag{E.2}
$$

$$
\begin{aligned}
P_1^{e_i} &\rightarrow r_{6,1}^{e_i} \\
P_2^{e_i} &\rightarrow r_{6,2|1}^{e_i} \\
P_3^{e_i} &\rightarrow r_{6,3|1,2}^{e_i} \\
P_4^{e_i} &\rightarrow r_{6,4|1,2,3}^{e_i} \\
P_5^{e_i} &\rightarrow r_{6,5|1,2,3,4}^{e_i}
\end{aligned}
$$

Notice that other probabilistic statements could be elicited in (E.2) according to the analysts preference. For example shorter conditioning sets might be considered: $P_1^{e_i} = P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_1 \geq v_{1,q_{50}}^{e_i}\right), \ldots, P_5^{e_i} = P\left(V_6 \geq v_{6,q_{50}}^{e_i} \mid V_5 \geq v_{5,q_{50}}^{e_i}\right)$. Another option would be to elicit a joint distribution $P_1^{e_i} = P\left(V_6 \geq v_{6,q_{50}}^{e_i}, V_1 \geq v_{1,q_{50}}^{e_i}\right), \ldots, P_5^{e_i} = P\left(V_6 \geq v_{6,q_{50}}^{e_i}, \ldots, V_5 \geq v_{5,q_{50}}^{e_i}\right)$ instead of a conditional probability of exceedance[11]. If the medians of each expert differ then the probabilities above are taken over different events and a correction needs to be computed. Notice also that the recommended choice for the percentile used in the probabilities stated is the median however any other percentile $v_{j,q_k}^{e_i}$ may be used. In particular other percentiles are necessary for discrete variables.

2. Take linear pooling of expert's assessments for V_i to determine $F_{V_j}^{DM}$ (The cumulative distribution function of the decision maker) and compute the median value $v_{j,q_{50}}^{DM}$ for the Decision Maker.

[11] Probabilities of concordance or discordance might be another option.

3. Compute the probabilities that each expert "*would have stated*" if he/she had been asked about the same quintile (median in the example above) as the Decision Maker (E.3) such that his/her estimates for the rank correlations remain unchanged. These probability will be denoted as $P_{1^*}^{e_i},...,P_{5^*}^{e_i}$

$$
\begin{aligned}
r_{6,1}^{e_i} &\rightarrow P_{1^*}^{e_i} \\
r_{6,2|1}^{e_i} &\rightarrow P_{2^*}^{e_i} \\
r_{6,3|1,2}^{e_i} &\rightarrow P_{3^*}^{e_i} \\
r_{6,4|1,2,3}^{e_i} &\rightarrow P_{4^*}^{e_i} \\
r_{6,5|1,2,3,4}^{e_i} &\rightarrow P_{5^*}^{e_i} \\
P_{1*}^{e_i} &= P\left(F_{V_6}^{e_i}(V_6) \geq F_{V_6}^{e_i}(v_{6,q_{50}}^{DM}) \mid F_{V_1}^{e_i}(V_1) \geq F_{V_1}^{e_i}(v_{1,q_{50}}^{DM})\right) \\
P_{2*}^{e_i} &= P\left(F_{V_6}^{e_i}(V_6) \geq F_{V_6}^{e_i}(v_{6,q_{50}}^{DM}) \mid F_{V_1}^{e_i}(V_1) \geq F_{V_1}^{e_i}(v_{1,q_{50}}^{DM}), F_{V_2}^{e_i}(V_2) \geq F_{V_2}^{e_i}(v_{2,q_{50}}^{DM})\right) \\
P_{3*}^{e_i} &= P\left(F_{V_6}^{e_i}(V_6) \geq F_{V_6}^{e_i}(v_{6,q_{50}}^{DM}) \mid F_{V_1}^{e_i}(V_1) \geq F_{V_1}^{e_i}(v_{1,q_{50}}^{DM}),...,F_{V_3}^{e_i}(V_3) \geq F_{V_3}^{e_i}(v_{3,q_{50}}^{DM})\right) \\
&. \\
&. \\
&. \\
P_{5*}^{e_i} &= P\left(F_{V_6}^{e_i}(V_6) \geq F_{V_6}^{e_i}(v_{6,q_{50}}^{DM}) \mid F_{V_1}^{e_i}(V_1) \geq F_{V_1}^{e_i}(v_{1,q_{50}}^{DM}),...,F_{V_5}^{e_i}(V_5) \geq F_{V_5}^{e_i}(v_{5,q_{50}}^{DM})\right)
\end{aligned}
\tag{E.3}
$$

An example of the situation where the expert's estimate and the Decision Maker's median differ is presented in Figure E.4. Observe that expert's e_1 median for V_1 is 50 while the DM's median for the same variable is 100. In this case the DM's median realizes the 66% percentile of expert e_1 distribution.

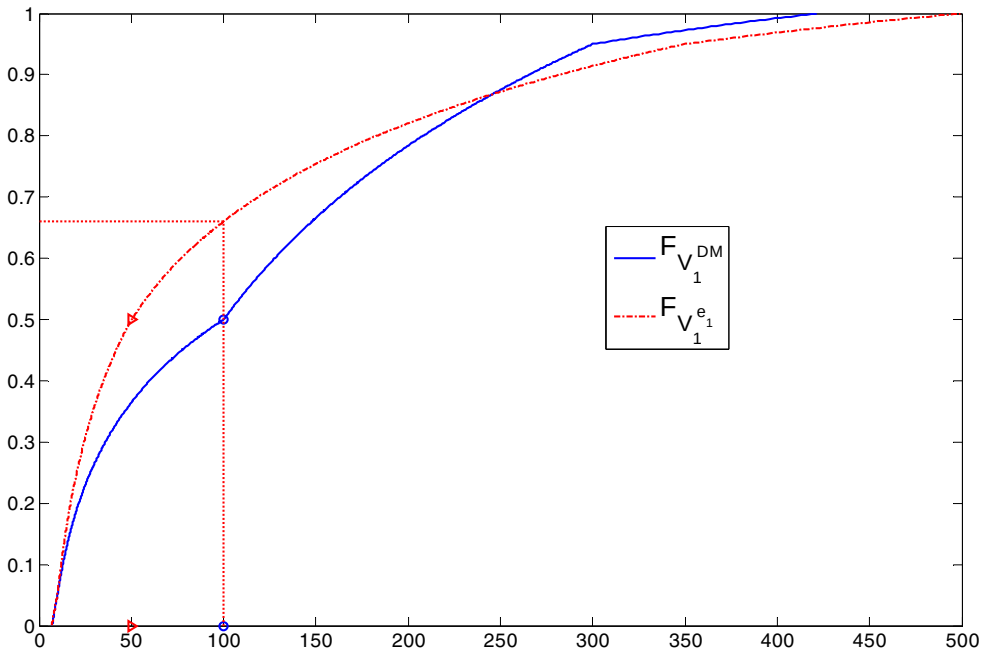


Figure E.4: Expert 1 and DM disagreement in the median for V1

Suppose for simplicity that both the DM and expert e_1 agree in the median value for variable V_6. Suppose that expert 1 answered:

$$\begin{aligned} P_1^{e_1} &= P\left(V_6 \geq v_{6,q_{50}}^{e_1} \mid V_1 \geq 50\right) \\ &= P\left(F_{V_6}^{e_1}(V_6) \geq 0.5 \mid F_{V_1}^{e_1}(V_1) \geq 0.5\right) = 0.75 \rightarrow r_{6,1} = 0.7 \end{aligned}$$

Then according to Figure E.5 $r_{6,1} = 0.7$. This may also be observed in Figure E.6 in the upper figure. Both V_1 and V_6 have been transformed to uniform and 10,000 samples with correlation 0.7 shown in a scatter plot. The expert's probability is approximately $P_1^{e_1} \approx \#\text{ of points in B}/\#\text{ of points in (B+D)}$. According to the method described above and Figure E.5, we have:

$$\begin{aligned} P_{1^*}^{e_1} &= P\left(V_6 \geq v_{6,q_{50}}^{e_1} \mid V_1 \geq 100\right) \\ &= P\left(F_{V_6}^{e_1}(V_6) \geq 0.5 \mid F_{V_1}^{e_1}(V_1) \geq 0.66\right) = 0.83 \rightarrow r_{6,1} = 0.7 \end{aligned}$$

In other words the expert 'would have stated' a probability of 0.83 for his/her estimate of dependence to remain constant if asked about the percentile corresponding to the Decision Maker's median. This may be observed in Figure E.6 in the lower figure. As before, V_1 and V_6 have been transformed to uniform and the same 10,000 samples with correlation 0.7 shown in a scatter plot. Again

$P_{1^*}^{e_1} \approx \#\text{ of points in } B_1 / \#\text{ of points in } (B_1 + D_1)$. A similar kind of correction needs to be applied to all probabilistic statements of all experts in relation (E.2) by applying relation (E.3).

4. Take linear pooling (E.4) of $P_{j*}^{e_i}$ i = 1, … , N for all j to determine (E.5) as explained in Morales et al [2007].

$$P_j^{DM} = \sum_i w_{e_i} P_{j^*}^{e_i} \tag{E.4}$$

$$\begin{aligned} P_1^{DM} &\rightarrow r_{6,1}^{DM} \\ P_2^{DM} &\rightarrow r_{6,2|1}^{DM} \\ P_3^{DM} &\rightarrow r_{6,3|1,2}^{DM} \\ P_4^{DM} &\rightarrow r_{6,4|1,2,3}^{DM} \\ P_5^{DM} &\rightarrow r_{6,5|1,2,3,4}^{DM} \end{aligned} \tag{E.5}$$

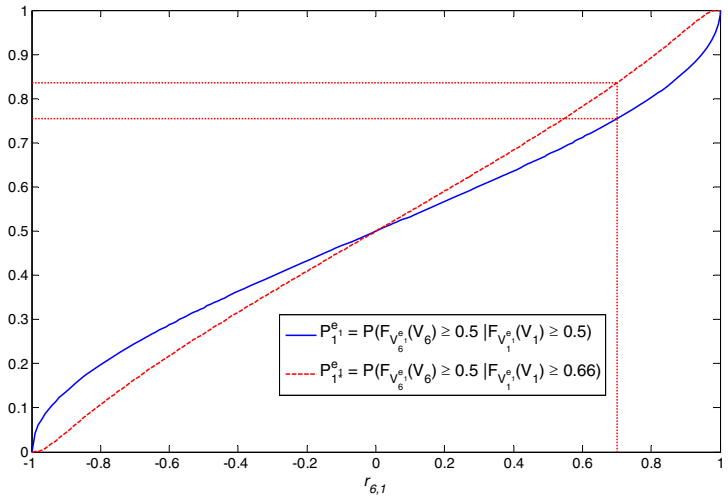


Figure E.5: Expert 1 and DM disagreement in the median for V1

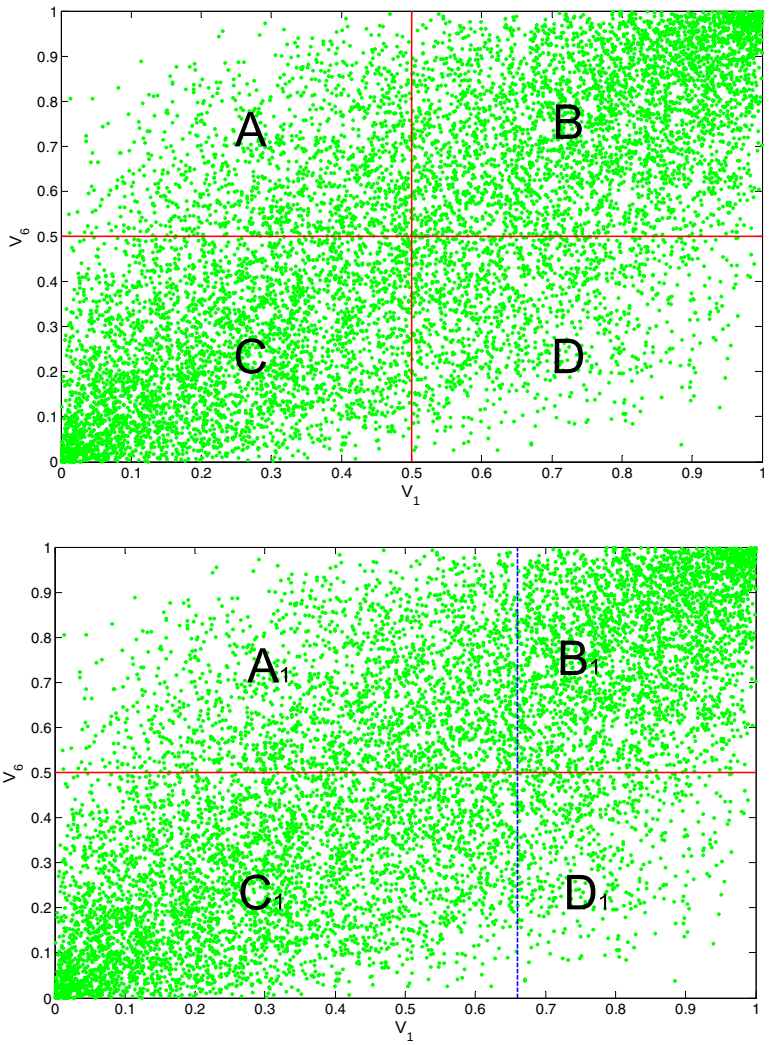


Figure E.6: Expert 1 and DM disagreement in the median for V1 (rank correlation 0.7)