



Episode 3
D2.4.3-03 - Note On Risk Model Validation

Version : 1.01

EPISODE 3

Single European Sky Implementation support through Validation



Document information


Programme	Sixth framework programme Priority 1.4 Aeronautics and Space
Project title	Episode 3
Project N°	037106
Project Coordinator	EUROCONTROL Experimental Centre
Deliverable Name	Note On Risk Model Validation
Deliverable ID	D2.4.3-03
Version	1.01

Owner

Eric Perrin EUROCONTROL

Contributing partners

AENA, CAST, DFS, ERC, INECO, NATS, NLR

	Episode 3 D2.4.3-03 - Note On Risk Model Validation	<i>Version : 1.01</i>
---	--	-----------------------

DOCUMENT CONTROL

Approval

Role	Organisation	Name
Document owner	EUROCONTROL	Eric Perrin
Technical approver	EUROCONTROL	Giuseppe Murgese
Quality approver	EUROCONTROL	Ludovic Legros
Project coordinator	EUROCONTROL	Philippe Leplae

Version history

Version	Date	Status	Author(s)	Justification - Could be a reference to a review form or a comment sheet
1.00	23/09/2009	Approved	Eric Perrin	Approved by the EP3 project consortium.
1.01	24/09/2009	Approved	Catherine Palazo	Minor format changes



TABLE OF CONTENTS

EXECUTIVE SUMMARY	4
1. INTRODUCTION	6
2. DEFINITION OF VALIDATION	7
3. VALIDATION STRATEGY	8
4. VALIDATION STATUS	9
4.1 VALIDATION OF MODEL DEVELOPMENT PROCESS	9
4.1.1 <i>Specification</i>	9
4.1.2 <i>A Research Delivery Model</i>	10
4.1.3 <i>Resources</i>	11
4.1.4 <i>Competence</i>	12
4.1.5 <i>Technical Quality</i>	12
4.1.6 <i>Compliance</i>	12
4.1.7 <i>Verification</i>	14
4.1.8 <i>Peer Review</i>	14
4.2 VALIDATION OF MODEL USABILITY.....	14
4.2.1 <i>Sensitivities</i>	14
4.2.2 <i>Confidence Ranges</i>	15
4.2.3 <i>Ranges of Validity</i>	15
4.3 VALIDATION OF MODEL RESULTS	16
4.3.1 <i>Calibration</i>	16
4.3.2 <i>Empirical Validation</i>	17
4.3.3 <i>Convergent Validity</i>	17
4.3.4 <i>Achievability</i>	20
4.3.5 <i>Face Validity</i>	20
5. CONCLUSIONS	22
6. REFERENCES	24
7. GLOSSARY OF TERMS	25
8. ANNEX - UNCERTAINTIES IN THE IRP	26
8.1 UNCERTAINTY IN EVENT FREQUENCIES DUE TO DATA QUANTITY	26
8.2 UNCERTAINTY IN EVENT PROBABILITIES DUE TO DATA QUANTITY	26

LIST OF FIGURES

Figure 1: Research Delivery Model.....	10
Figure 2: Calibration Against Trend in Fatal Accident Frequencies since 1990.....	16



EXECUTIVE SUMMARY

The Integrated Risk Picture (IRP) is a key tool used in the modelling of the accident risks in the SESAR Concept of Operations. It is desirable to explain how far the IRP can be considered to be validated.

The validation strategy for IRP includes the following elements:

- Validation of the model development process – ensuring that the model has been developed in a sound, defensible, well-grounded way.
- Validation of the model usability – ensuring that users of the model receive adequate guidance about its validity.
- Validation of the model results – ensuring that the risk picture results are trustworthy.

The current status on each element is as follows:

- The model development process has been fully validated by ensuring adequacy in the following aspects:
 - Specification for the model
 - Resources for model development
 - Competence of the model developers
 - Technical quality of the model
 - Compliance with the specification
 - Verification of the model construction
- The usability of the model is being validated through on-going definition of:
 - Ranges of validity for model inputs
 - Sensitivities of results to model inputs and parameters
 - Confidence ranges in the results
- The model results have been validated to the extent that is practical at present through:
 - Calibration against ATM changes during the period 1990-2005 (to be updated - see Recommendation #3 hereafter)
 - Empirical validation against independent estimates of ATM overall contribution
 - Convergent validity against available statistics on ATM-related incident rates
 - Face validity through acceptance of the model by a limited group of stakeholders

In future work, further validation would be desirable. It is recommended that this should include:

Recommendation # 1: Peer review for IRP with progressively wider groups of stakeholders. This should be facilitated by the production of two types of report: (i) analytical report with details of the model; and (ii) operational reports containing the results and recommendations.

Recommendation # 2: Explore the modelling of user inputs ranges in the IRP further development.



Episode 3
D2.4.3-03 - Note On Risk Model Validation

Version : 1.01

Recommendation # 3: Modelling in the IRP the complete set of ATM changes that have taken place since 2005 and update the calibration of the IRP model against historical trends.

Recommendation # 4: Progressive calibration against accident and incident experience as it accumulates.

Recommendation # 5: Calibration of risks for specific regions or units for which there is suitable accident or incident data.

Recommendation # 6: Further collection of alternative parameter estimates for use in defining confidence ranges.

Recommendation # 7: Documentation of convergent validity as accident and incident datasets are enlarged.

Recommendation # 8: Review IRP initial predictions for SESAR with wider groups of experts in the subject, and ensure consistency between the IRP and the appropriate safety case. This may require a tailor-made training on the IRP (or equivalent accident-incident model) to ensure that each participant understands the approach to risk modelling.

Recommendation # 9: Collection of in-service performance and actual design targets for ATM elements, in order to check achievability of safety requirements

Recommendation # 10: Review of the face validity of IRP and its conclusions with progressively wider groups of stakeholders during the SESAR Development Phase.



1. INTRODUCTION

The SESAR Top-Down Systemic Risk Assessment (**Ref 1**) explains how the Integrated Risk Picture (IRP) can be used to quantify the accident risks in the SESAR Concept of Operations, thereby supporting a systemic risk assessment of the SESAR. It:

- Explains how IRP has been applied as a prototype to model SESAR (first evaluation), identifying the main uncertainties and information gaps that would need to be addressed to complete the work.
- Explains what types of results are available from IRP, outlining their potential utility for the safety management of SESAR. This includes recent improvements in the presentation of positive contributions of ATM elements (OI steps) to accident risks.
- Presents the preliminary estimates made by IRP of the accident risks in the SESAR Concept of Operations (an initial SESAR Risk Picture), showing whether it achieves the overall SESAR safety target. The work is intended to deliver consistent design targets for SESAR concept components, which together will be sufficient to achieve the overall SESAR safety target.

This note explains how far the Integrated Risk Picture (IRP) can be considered to be validated. It considers various possible meanings of “validation”, and outlines a validation strategy based on a combination of all of them. It then explains the extent to which the IRP has been validated by work to date, and recommends further work where necessary. This document should be read in conjunction with the SESAR Top-Down Systemic Risk Assessment (**Ref 1**) since it contains all relevant figures that are referred herein. In addition, full details on IRP itself are provided in section 11 of (**Ref 1**).



2. DEFINITION OF VALIDATION

In general, anything that is “valid” is sound, defensible and well-grounded. In the case of a model such as IRP, valid also means that it meets its specification and produces the required results. In addition, one way of evaluating whether the model is fit for purpose might be to consider whether it is capable of leading the user towards sensible decisions about risk reduction. To “validate” something is to check or ensure that it is valid. “Validation” of IRP is therefore the process of checking that it is sound, defensible, well-grounded, meets its specification and produces the required results (including the credibility aspect).

On a complex model, such as IRP, there are many different types of check that could be performed to show whether it is valid in different respects. These different checks account for the many different interpretations of the term “validate”. Comprehensive validation ideally requires many different checks, covering the complete range of possibilities.

Because the model can produce an almost unlimited range of possible results, there is no absolute limit to the validation work that could be carried out. To date, validation has been pursued in each respect to the extent that was judged appropriate for the stage of development of the model. In future work, each type of validation could be extended. This implies that validation is never complete, as more work can always be done. As a result, the point at which the model can be described as “validated” is necessarily chosen by judgement rather than objectively defined.



3. VALIDATION STRATEGY

The Integrated Risk Picture (IRP) is a key tool used in the modelling of the accident risks in the SESAR Concept of Operations.

The validation strategy for IRP includes the following elements:

Validation of the model development process

- Specification - ensuring that the specification on which the model is based reflects the needs of the users.
- Resources - ensuring that the resource for model development is appropriate for the demands implied by the specification.
- Competence - ensuring that the model is developed by people with appropriate skills.
- Technical quality - ensuring that the model achieves an appropriate balance between technical sophistication and practicality.
- Verification - checking that the model was constructed “correctly”, and contains no unintended errors.
- Peer review - checking that the model has an appropriate degree of scientific rigour, as judged by independent experts.
- Compliance - checking that the model meets its specification and delivers the required results.

Validation of the model usability

- Sensitivities - ensuring that the relationships between user inputs, model parameters and results are fully understood.
- Confidence ranges - showing the ranges of uncertainty attached to the model results.
- Ranges of validity - showing the ranges of user inputs that the model is able to accept.
- Validation of the model results (inc. current ATM-related risks where data is available and predictions of how those risks will change in response to the planned changes to the ATM system).
- Calibration - checking that the model correctly predicts the effects of previous interventions whose impact is known.
- Empirical validity - checking that the model gives results that are the same as other, independent models, to within their confidence limits.
- Convergent validity - checking that the model gives results that are consistent with other results, where convergence would be expected as the models are improved.
- Achievability - checking that safety requirements derived from the model are reasonably practicable to achieve in practice.
- Face validity - the on-going process of stakeholder acceptance of the model.



4. VALIDATION STATUS

The following sections outline the status on each element of the validation strategy detailed in Section 3.

4.1 VALIDATION OF MODEL DEVELOPMENT PROCESS

4.1.1 Specification

The requirement for an integrated risk picture for ATM in Europe was set by the EUROCONTROL High Level European Action Group for ATM Safety (AGAS) as one of the priority actions (High Priority Action Area 8) to improve safety in European airspace (**Ref. 2**).

The project to develop the IRP was initiated by EUROCONTROL, working with the EUROCONTROL System View Cell (SVC), the three Research Areas (SSP- Sector Safety and Productivity; NCD - Network Capacity and Demand; APT - Airport Throughput), DAS/SSM (Safety and Security Management) and DAP/SAF (Safety Enhancement); then part of DAP/SSH (Safety, Security and Human Factors) and now part of Safety, Security and Human Factors in ATM Network Support & Services Area. The specification for the tool was driven by the need to provide answers to key questions that were:

- What is the safety assessment of the overall system?
- How might these new elements interact?
- Are there negative interactions that can be avoided, or even positive interactions, as yet unplanned into the system design concept, which could yield extra safety?
- Where are the strong and weak safety areas in the overall system?
- Is the resultant system risk sensitive to the sequence and timing of implementation?
- During the implementation phase, how do we demonstrate we are still on target?
- What happens if expected safety impacts of ATM changes fall short?

The project was coordinated with the Federal Aviation Administration (FAA) through the FAA/EUROCONTROL Action Plan 15 (**Ref. 3**), which required a preliminary model of ATM safety, highlighting important interactions and dependencies between system elements, showing the relative risks and where safety needed to be improved.

The specification for the IRP was developed by EUROCONTROL in 2003, based on long experience of the needs of ATM for comprehensive risk models. In subsequent years, the specification has been extended in response to model development and stakeholder reactions in EUROCONTROL, and the FAA.

The validity of the IRP specification is reinforced by the fact that it is similar to the specification of the Causal Model of Air Transport Safety (CATS)¹ project subsequently commissioned by the Netherlands Government in 2004.

During the SESAR Definition Phase, the performance analysis of the SESAR ATM Target Concept in the safety focus area (**Ref. 4**) has identified the need for *“an agreed model of incident-accident relationship, to be used as the basis for an Operational safety Focus area*

¹ The Causal Model of Air Transport Safety (CATS) has been developed by a consortium including Delft University of Technology (TUD), National Aerospace Laboratory (NLR) and White Queen (WQ) in The Netherlands, and Det Norske Veritas (DNV) in the UK. The motivation for the project is the need for a thorough understanding of the causal factors underlying the risks of air transport so that efforts to improve safety can be made as effective as possible. The project was commissioned by the Dutch Ministry of Transport.



decomposition in the Performance Framework” in order to decompose the observed frequency of ATM-attributable accidents into lower level targets (SESAR D3 p63). After screening of the Target Concept for impacts on safety regulation, it was recommended that “a new accident model should be developed that represents the SESAR operational concept (related to redefinition of ATM scope functions and boundaries)” (SESAR D3 p87).

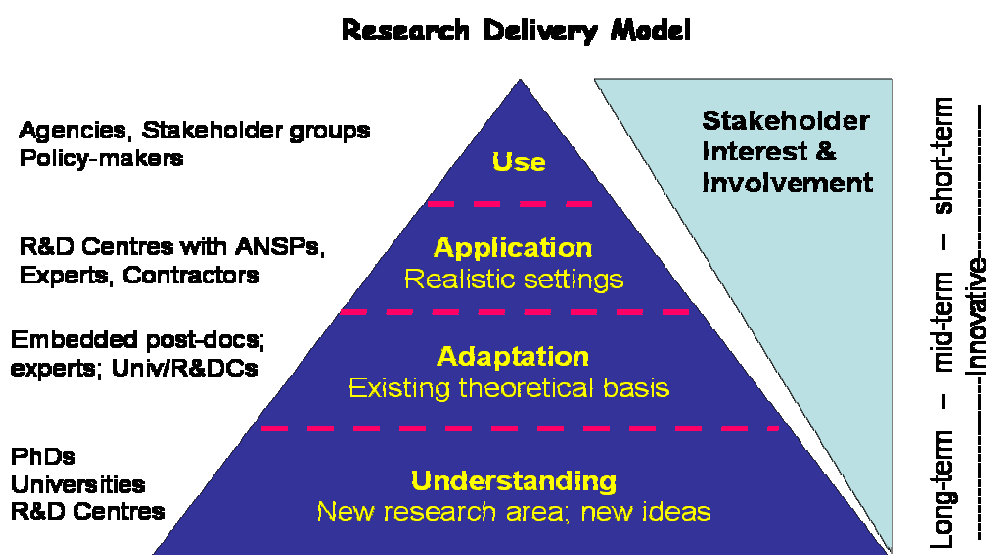
EP3 has used the IRP to model the accident risks in the SESAR Concept of Operations, to show how it can achieve its target for improving ATM safety and to apportion the target between the constituent components of the ATM system. In EP3 WP2.4.3, it has been further developed to represent each of the operational improvements (OIs) in the SESAR Concept of Operations and provide a first quantitative estimate of their overall effects on safety, so as to predict whether SESAR will be able to meet its quantitative safety target.

The Safety Roadmap in SESAR D5 (Ref. 5) and SESAR D6 (Ref. 6) for Service Levels 0&1 preparation identify the need to develop a SESAR-compliant accident-incident model and SESAR Safety Target Achievement Roadmap (STAR). Building upon (Ref. 2), the SESAR accident-incident model and STAR should be further developed in SESAR SWP16.1 (Ref. 7).

4.1.2 A Research Delivery Model

Research is needed when there is a desire to improve, and when going through a significant period of change. Both of these cases currently apply to ATM, whether driven by short term individual ANSP goals, or by larger and longer term collective programmes such as SESAR in Europe or NextGen in the US, which aim to reach the next level of evolution in ATM and deliver the increased growth and quality of service demanded by customers. A model of doing research and getting it into practice is needed – otherwise many safety people would say ‘why bother? Spend the money elsewhere!’ Therefore a model is needed to help ensure that safety research ultimately reaches its goal, of making real systems safer. This model is called the research delivery model and is shown in Figure 1.

Figure 1: Research Delivery Model



It highlights four levels of research: understanding, adaptation, application, and usage. These typically correspond to PhD, Post-doctoral study, expert application (e.g. internal expert or contractor studies), and collaborative studies between EUROCONTROL Agency staff and stakeholders. The latter level also includes taking the research results into the industry, whether via stakeholder consultation groups, guidance material for ANSPs, or via formal common implementation processes which ensure that useful research results can be spread across the ECAC states.



The IRP development has followed the Research Delivery Model as follows:

- **Understanding in 2003/2004:** specific efforts were dedicated to judge the appropriate level of research required relying on sharing knowledge about what has been done, what is being done in terms of risk modelling, where the 'hot' areas in Safety R&D are, and where the dead-ends are. This required collaboration on Safety R&D, and a network for such research players in the ATM domain to share knowledge and experiences (e.g. within the framework of Action Plan 15).
- **Adaptation in 2005/2006:** the IRP risk modelling has been improved over years in the areas of target level of safety apportionment, safety impacts of ATM (including positive contributions), and construction of a safety roadmap for future ATM changes. In addition, the source of the data (pedigree) has been made plain, along with the collected results of sensitivity analyses, and the degree of uncertainty in the results.
- **Application in 2006-2009:** In order to maximize the chance of the IRP to 'make-it' into the real world:
 - Causes of the accidents and incidents, and the historical influences on them, have been combined to form the risk picture for 2005 (baseline risk picture).
 - ATM changes that are expected to be in place by 2012 have been modelled to form the risk picture for 2012 (**Ref. 8**).

Development above was a good starting place for risk modelling in EP3, and consequently:

- Within the framework of EP3 (**Ref. 1**):
 - The baseline risk picture has been updated, and
 - 3 first estimates of changes in risk in 2020 have been developed, namely: (i) risk that would result if no changes were made to ATM safety; (ii) future risk picture with SESAR implemented; and (iii) indicative risk picture complying with the SESAR Safety Target.
- **Use:** It is planned that future development of such risk model would take place during the SESAR Development Phase. The first three levels of research described above has helped 'grade' stakeholders' expectations appropriately, and enabled them to see which R&D might help them, and then to join in the process of making it succeed in the SESAR Development Phase. To support the transfer to usage, a training course to help individual users understand and use the models has been developed (**Ref. 9**). However, it would be very time-consuming to train all potential users in this way, a computer-based training (CBT) has also been developed (**Ref. 10**), which is expected to be a more efficient method of training large groups of users.

4.1.3 Resources

The required resources for the IRP development have been secured each year since 2004. IRP development has been one of EUROCONTROL's largest safety research projects during this period. For the current activities in EP3 on risk modelling activities and validation, allocated resources have been appropriate to develop the initial SESAR risk picture and to provide a useful trial of the methods.

Refinement of the model will be completed within the SESAR JU Programme in WP16.1.1.1.

The adequacy of the resources is demonstrated by the rapid progress made by IRP as illustrated in section 4.1.2.



4.1.4 Competence

Technical development of the IRP has been performed by safety professionals considered competent for the task due to their familiarity with the modelling techniques and the historical data. The core competencies were: Systems thinking, hazard analysis, accident/incident investigation and safety verification. The safety professionals system have in particular applied a planned, disciplined, systematically organized and before the fact process to evaluate all elements and interdependencies of the system. They have also brought subject matter experts (CNS/ATM) and domain experts (ops, eng, maintenance) to the table, although only a few brief sessions with domain experts to estimate parameters have been used so far.

4.1.5 Technical Quality

The technical work underlying IRP is fully documented in (**Ref. 1**) and a Methodology Report (**Ref. 11**). The latter explains the reasons for the choice of the particular modelling approach, which was based on a review of previous approaches to similar requirements. The chosen methodology includes a detailed analysis of a substantial dataset of accidents and incidents, a complex model based on fault tree and influence models, which is founded on an explicit model of the ATM process, and delivers risk results for user-selected cases. The roadmap development (**Ref. 12**) allows optional selection from the major ATM changes planned to occur in the period to 2020. The model represents the “state of the art” in the field.

The suitability of the fault tree approach is indicated by the fact that the technique has been extensively used in aviation and many other industries, and has been found appropriate for safety assessment work in EUROCONTROL over the last 10 years. Known weaknesses of fault trees compared to networks have been addressed at this stage of development as follows:

- Causes that cannot be split into simple failed/operating states are represented through the influence model.
- Top-down quantification is used for the baseline risk model to ensure that the risks are properly calibrated against actual accident experience.
- Errors arising from lack of independence between the base events are minimised by selection of barriers that are as independent as possible, and representation of common cause events through the influence model.

In the longer term, more substantial improvements could be considered by modelling the influences through networks. A Bayesian network should make the influence model implementation more robust. This has been described in Annex III of (**Ref. 1**).

The technical quality of the work is also indicated by the fact that similar models are being adopted by the CATS project and the FAA (**Ref. 3**).

4.1.6 Compliance

4.1.6.1 Evaluation against results requirements

Based on (**Ref. 1**), the evaluation of the presented methodology is that it meets the requirements for the IRP as detailed in the EP3 Description of Work (DOW) (**Ref. 14**) as the quantitative risk model outlined above provides a detailed representation of the role of ATM in accident causation:

- The overall contribution of ATM to aviation risk and the relative importance of different accident categories are clearly shown in results such as Table 1 in (**Ref. 1**).
- The causal factors underlying the ATM contribution are shown in results such as Figure 10 in (**Ref. 1**).



- The contribution of ATM in both causing and preventing aviation accidents is shown in results such as Figure 11 in (Ref. 1).
- (Ref. 2) makes clear the degree of uncertainty in the results, and although it is not practical to show this information succinctly for all results, all necessary information is shown for selected results. This is shown in results such as Figure 7 in (Ref. 1).
- Consistent with SESAR D4 (Ref. 15) which makes the following key points concerning ATM safety:

*“...the need for ATM to maximize its contribution to aviation safety and
the need for ATM to minimize its contribution to the risk of an accident”*

the existing benefits of ATM in preventing accidents is shown in results such as Figure 9 in (Ref. 1).

4.1.6.2 Evaluation against defined user cases

It meets the full set of defined user cases as follows:

- Strategic direction for safety improvements – this refers to the strategic direction of safety improvements or safety research. Users may be concerned with ATM as a whole, or with particular research domains. The requirement is to understand the priorities or key safety issues facing ATM (or their part of ATM). In other words, they need to know the most likely causes through which ATM may be involved in future aviation accidents. This is illustrated in results such as section 6.10 of (Ref. 1).
- Safety impacts of individual ATM changes – this refers to the effects of individual projects that change ATM in some way. Users may be conducting safety assessments, or evaluating the need for such studies. Their requirement is to understand the main ways in which their project may affect ATM safety. This includes both potential benefits that should be maximised, as well as hazards (potential negative impacts) that must be minimised. This is illustrated in results such as Table 13 in (Ref. 1).
- Overall safety target compliance – this refers to compliance of the ATM system as a whole with overall safety targets. The requirement is to predict whether the planned individual ATM changes, once implemented, will combine to deliver the required improvement in overall ATM safety. This must take account of interactions between the ATM changes, as well as expected traffic increases and other changes in the aviation system. This is illustrated in results such as Figure 17 of (Ref. 1).
- Safety target apportionment – this refers to appropriately apportioning overall safety targets to provide safety objectives for individual systems. The requirement is to apportion safety targets in a way that maximises safety improvements, while also being realistic about practical achievements. The apportionment must also take account of potential interactions between individual systems, which may reduce the expected safety benefits once in service. This is illustrated in results such as Table 19 of (Ref. 1).
- Risk picture for specific units – this refers to obtaining a risk picture for specific units. A “unit” may be an airport, TMA airspace or en-route sector. The unit-specific risk picture may be used to support any of the other user cases applied to that specific unit. The methodology is explained in (Ref. 13).
- Safety roadmap – this refers to definition of the sequence of changes between the present and the planned future ATM system, so that the safety target is met at all stages, and in particular that risks are decreased where possible. This is illustrated in results such as Figure 14 of (Ref. 1). The methodology has been implemented in the form of an Excel tool (i.e. the STAR version of IRP) (Ref. 16).



- Alignment of severity classifications - the alignment of severity classification schemes that are used in various regulatory requirements, safety assessments and data collection and analysis schemes. The IRP enables to get an understanding of the remaining safety margin between the end state and the accident scenario. This is illustrated in results such as Figures 8 and 9 of (Ref. 1).
- Safety performance monitoring – this refers to the monitoring of safety performance. Users may be responsible for monitoring at specific units (e.g. airspaces, airports etc) or for long-term tracking of safety performance to verify compliance with the safety roadmap. At present, trend monitoring is done for some incident types, but not consistently for all safety-critical ones. This is illustrated in results such as Table 19 of (Ref. 1).

4.1.7 Verification

Verification helps ensure that the model was constructed correctly, and contains no unintended errors. It involves:

- Independent review of each part of the work (see section 4.1.7 below).
- Full documentation of the work through project reports (see section 6).

Verification cannot guarantee that there are no errors in a model as complex as IRP. Nevertheless, a high degree of error correction is achieved, through use of the following self-checks:

- The fault tree model is implemented in a gate-by-gate form, showing all intermediate results, which are also included in the project reports. This allows manual checks of each stage in the model, and this has been effective in identifying errors in the model.
- The fault tree model is implemented twice, quantified once from the top down, and once from the bottom up. The fact that this returns numerically identical risks helps trap a high proportion of errors in model construction.
- The contributions of causal factors for each barrier in the fault tree sum to 1, due to the choice of a non-dimensional risk reduction metric, which allows a simple check against numerical errors in quantification of these results.

4.1.8 Peer Review

An additional check that the model has an appropriate degree of scientific rigour, as judged by independent experts, can be based on expert review or publication in refereed journals. This process is being used in the CATS project, and has led to greater concentration on dependencies, a subject already addressed in detail in IRP. A difficulty of obtaining peer review for IRP is the limited availability of ATM risk modelling experts.

Recommendation # 1: *Peer review for IRP with progressively wider groups of stakeholders. This should be facilitated by the production of two types of report: (i) analytical report with details of the model; and (ii) operational reports containing the results and recommendations.*

4.2 VALIDATION OF MODEL USABILITY

4.2.1 Sensitivities

The IRP model implements complex relationships between user inputs, model parameters and results. The model is structured to show the sensitivity of the results to the user inputs. However, clarifying the sensitivities of the results to the other model parameters is a major



challenge for the work. This is covered in the IRP results by the “contributions” of each model parameter. Because the model is non-linear, the contribution is rather complex. Two key types of contribution are distinguished (see (Ref.1) section 11.6.2):

- Negative contribution - the risk caused by unsuccessful ATM, representing the times when ATM failures (including technical failures, ineffective human performance or non-availability of equipment) have contributed to the causes of an accident. Risk reduction worth is a measure of negative contribution.
- Positive contribution - the existing benefits of ATM in preventing accidents. Risk achievement worth is a measure of positive contribution.

Risk reduction worth and risk achievement worth are included in sets of results. This is illustrated in results such as Figure 10 and Table 13 of (Ref. 1).

4.2.2 Confidence Ranges

All results could be expressed as probability distributions representing uncertainty about their true values. It shows how different values of the parameter are associated with different probabilities. Although informative, this presentation requires the most space, as each best-estimate value must be accompanied by a graph of uncertainties. In the case of IRP, which is intended to show the breakdown of contributors to risks, adding probability distributions to each contribution would greatly increase the size of the reports.

The benefit of the distributions is questionable, as they are rarely read in detail, and serve mainly as indicators that uncertainties have been taken fully into account. Hence this presentation is most suitable for the most important results (e.g. risk measures to be compared with safety targets).

The confidence range is the most common brief representation of uncertainties. It shows the values with given probabilities of exceedence. For simplicity, these are summarised as best-estimates (resulting from point estimates in the spreadsheet model) and confidence limits (the 5%ile and 95%ile from the Monte Carlo analysis of uncertainties).

For uncertainty due to data quantity, two cases are used in the IRP:

- For frequency data, events are assumed to follow a Poisson distribution. This is outlined in section 8.1.
- For probability data, confidence limits are obtained from the number of failures and demand using a beta distribution. This is defined in section 8.2.

Uncertainty due to data choices is defined using alternative data choices. The largest and smallest values from plausible alternative choices are used to define the confidence limits from this source. The uncertainty distribution is assumed to be triangular between these values, with a peak at the best estimate.

Where available, they are illustrated on plots using I-shaped bars where the upper ($p_{0.95}$) and lower ($p_{0.05}$) ends represent the upper and lower confidence limits respectively. This is shown in results such as Figure 7 in (Ref. 1).

4.2.3 Ranges of Validity

A model such as IRP can only deliver valid results in response to valid inputs. At present IRP does not define the ranges of user inputs that are acceptable. This has not yet been explored in the IRP development.

Recommendation # 2: *Explore the modelling of user inputs ranges in the IRP further development.*



4.3 VALIDATION OF MODEL RESULTS

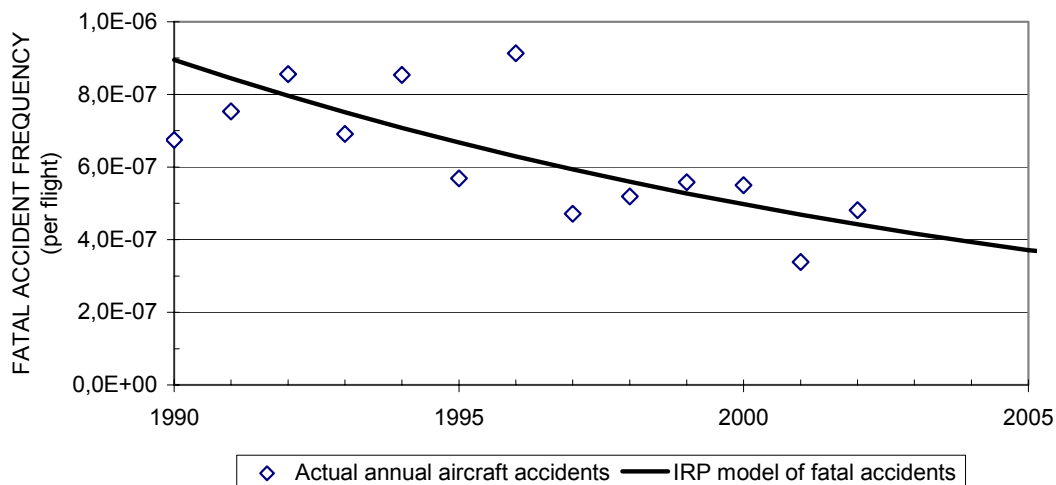
4.3.1 Calibration

Calibration involves checking that the model correctly predicts the effects of previous interventions whose impact is known. In aviation it is rarely possible to identify the effects of individual changes. Where this is possible, the effects have been used in the model, making it inappropriate to use these for independent calibration.

The IRP has therefore been calibrated against the combination of all changes that have occurred over the period 1990-2005. The changes to air traffic, the operating environment and to ATM itself that have occurred since 1990 have been defined, and represented through the inputs to the IRP model, in order to form a prediction of the risk levels and causal breakdown for average commercial flights in Europe in 1990, which can be compared with historical data.

For example, Figure 2 shows the calibration against frequencies for all types of fatal accidents on large Western jets world-wide. The historical values for each year since 1990 are compared with the IRP results for 1990 and 2005, which are connected with a smooth curve. In future work, it would be possible to model the specific time periods of the various changes, so as to construct a more realistic historical safety staircase.

Figure 2: Calibration Against Trend in Fatal Accident Frequencies since 1990



Although the historical data shows a substantial year-on-year variation, due to the relatively small numbers of events, it is clear that the modelled trend is consistent with it. This is not an independent empirical validation because the historical data is also used in developing the IRP. Nevertheless, it does give encouragement that the model is valid for future predictions.

Further calibration of this type is recommended:

- Validation against accident and incident experience in the coming years. For the precursor incidents, this would provide continuous validation, as well as contributing to improving the model.
- Retrospective prediction of risks for specific regions or units (based on (Ref. 13)) for which there is accident or incident data suitable for validation of aspects of the risk model. Accident datasets are available for different aircraft types and geographical regions. Incident data could also be used for specific airports or airspaces.

This historical calibration has not been updated since IRP 2005. It would be desirable to complete this exercise to improve the predictions for SESAR. The predictive mode in IRP



could be used to estimate the effects of ongoing ATM changes. This would provide a continuous calibration of the model. However, few of these ongoing changes are represented in the SESAR OIs, and therefore it would be necessary to make a detailed definition of these before attempting to model their effects (e.g. downlink of Selected Flight Level).

Recommendation # 3: *Modelling in the IRP the complete set of ATM changes that have taken place since 2005 and update the calibration of the IRP model against historical trends².*

Recommendation # 4: *Progressive calibration against accident and incident experience as it accumulates.*

Recommendation # 5: *Calibration of risks for specific regions or units for which there is suitable accident or incident data.*

4.3.2 Empirical Validation

Empirical validation in general involves comparing the model results with independent estimates, and demonstrating agreement to within the confidence range. This “classical” approach is impractical at present for most of the IRP results, because there are no directly comparable models and most available data has been used in constructing the IRP.

An exception is for the SRC Document 2 (“Aircraft Accidents/Incidents and ATM Contribution”) (**Ref 17**). Previous estimates of the contribution of ATM to aviation accidents have been in the range 2-6% (**Ref 17**). The results in (**Ref. 1**) show values ranging from 0.2% to 22%, depending on the definition of ATM contribution and the accident severity included. IRP estimates that only 1% of accidents are directly caused by ATM, through errors of commission by controllers or critical failures of their equipment. Given the range of values quoted by SRC, it is considered to validate this result from the IRP. However, no simple comparison is possible without further study of the definitions.

Validation using this approach becomes more difficult for more detailed results, because:

- Most available data has been used in constructing the model. When new data becomes available, this is usually because it has been collected specifically to improve resolution, with the intention of adding it to the IRP. It would be inappropriate to maintain this as an independent validation dataset outside the IRP. A possible use of this type of data is considered under convergent validation below.
- Some alternative data sources were discarded while constructing the model, often due to being less well defined than the chosen data. It would be inappropriate to use these sources as to validate the IRP, as they would inevitably conflict with it.
- No other model is currently able to obtain comparable results. Independent results imply different causal models, which makes simple comparison of the causal factors impractical, since their meaning often differs even when the names are the same. A possible use of this type of data is considered under face validity below.

Recommendation # 6: *Further collection of alternative parameter estimates for use in defining confidence ranges.*

4.3.3 Convergent Validity

Convergent validity in general involves checking that the model gives results that are similar to (or correlated with) other results, where convergence would be expected as the models are improved. This is a less demanding requirement than empirical validity, and may be more appropriate for IRP.

² In line with recommendations in Section 8.1 of (**Ref. 1**).



Convergent validity is a concept that can make use of new data such as this to progressively validate and improve the model. New data is regularly incorporated in IRP. The effects of the new data on the model can be considered a type of validation. If the new data produces only a small change in the model, then the model has more convergent validity. This validity can be measured using the changes in the results. If 90% of changes due to new data are within the 90% confidence ranges on the results, then these results could be considered validated, at least with respect to the new data.

There are several difficulties with this type of validation:

- There are many different possible results from IRP, and most new data affects only a few of them. It is not possible to select a single metric to measure all effects, without producing the misleading impression that most new data makes no difference to the results.
- The new data may show that changes are necessary in the model structure or parameter definitions. This will prevent direct comparisons before and after the new data in any parts of the model that have changed.
- In contrast to empirical validation, it is not necessary to preserve the independence of the new data. In fact, it is appropriate for the new data to be incorporated in the model, recording only the changes that occur in the results. The new data will also affect the confidence ranges:
- Additional accidents and incidents, added to improve the causal distribution data that underlie the fault tree models, will reduce the uncertainty due to data quantity.
- Alternative sources of event probabilities, replacing assumptions or preliminary data in the fault tree models, may reduce the uncertainty due to data choices, if they allow some previous data choices to be discarded as incorrect. Otherwise, they may give better estimates of the uncertainty, which may be larger or smaller.

These revised model results and uncertainty ranges will then become the basis for the convergent validation the next time data is added.

It is recommended that in future work involving new data for IRP, the convergent validity is checked and documented.

Recommendation # 7: *Documentation of convergent validity as accident and incident datasets are enlarged.*

4.3.3.1 Baseline Risk Picture

For the baseline risk picture, simple checks of validity of the overall results are made below:

- Fatal mid-air collisions involving commercial aircraft in Europe last occurred at Überlingen in 2002 and prior to that in Zagreb in 1976. Given the recent traffic growth since then, this is roughly consistent with the prediction from Table 5 of (**Ref. 1**) of an average return period of 8 years for this event.
- Fatal runway collisions involving commercial aircraft in Europe last occurred at Milan Linate in 2001 and Paris Charles de Gaulle in 2000. This is roughly consistent with the prediction from Table 5 of (**Ref. 1**) of an average return period of 5 years for this event.
- Fatal CFITs involving commercial aircraft in Europe last occurred at Isparta in 2007, Brest in 2003, Diyarbakir in 2003 and Zurich in 2001. Given the adoption of TAWS during this period, this is roughly consistent with the prediction from Table 5 of (**Ref. 1**) of an average return period of 3 years for this event.
- There have been no fatal wake turbulence accidents or taxiway collisions involving commercial aircraft in Europe. This is consistent with the prediction from Table 5 of



(Ref. 1) of an average return period of 30 and 200 years respectively for these events.

To validate the results above, it is desirable to compare against independent analyses. However, genuinely comparable and independent data sources are difficult to find, as the following examples show.

The SRC Annual Report (Ref 18) includes numbers of accidents on aircraft above 2250kg MTOW in ECAC. In 2005, it reported no mid-air collisions, 7 aircraft-aircraft collisions on the ground, and 5 CFIT accidents. However, this includes VFR aircraft, which increases the numbers and prevents a direct comparison with IRP.

Data from the Aviation Safety Network (Ref 19) shows an average of 6.7 fatal accidents per year in Europe for the 10 years to 2005. This is based on multi-engine airliners, but includes accidents in Russia, which is outside ECAC, and on fire-fighting aircraft (which together amounted to 35% of accidents in Europe during 2002-05). The data trend suggests that the number in 2005 is approximately 70% of that in 1990-2005. Hence the number of fatal accidents in 2005 could be estimated from this data as approximately 3.0 fatal accidents per year. This is consistent with the value of 3.2 estimated in Table 5 of (Ref. 1).

This is not an independent empirical validation because the historical data is used in developing the IRP. Nevertheless it does give encouragement that the model is valid for future predictions.

It is concluded validation of this type of risk model is difficult, but the baseline risk estimates are considered to be validated as far as practicable with the currently available information.

4.3.3.2 "Do Nothing" Risk Picture

To validate the model in the "do nothing" case, it would be desirable to obtain data that showed the effect on collision risks of changing the number of flights. For example, does the number of accidents really vary according to the square of the number of flights if nothing else changes? It might be expected that the accident frequencies per flight would be independent of traffic, and hence that the numbers of accidents would simply increase in proportion to the traffic.

In investigating this, it is difficult to obtain sufficiently large datasets with significant traffic changes independent of safety changes. One problem is there have been too few collisions of commercial aircraft to obtain reliable evidence. Trends have been obtained for frequencies of accidents of all types and for mid-air separation infringements since 1990, but these show the accident frequencies have remained roughly constant (Ref 20). If no other changes had occurred, this would be sufficient to invalidate the above assumption. In reality, it is believed that the safety improvements over this period have been just sufficient to offset any adverse effects from the increase in traffic. They have kept the numbers of accidents broadly constant. In the current model, which explicitly represents planned safety measures, the continuance of this balance cannot be taken for granted, and should be the result of the model not an input to it.

The available data does show what changes in accident patterns would have resulted in the absence of these safety improvements. In future work, it might be possible to segregate the data by time, day or season, in search of cases where traffic differs while safety measures remain constant. Meanwhile, the model above remains not validated. This is reflected in the wide confidence limits adopted above.

4.3.3.3 Initial SESAR Risk Picture

It is impossible at present to compare the above predictions with any independent source (no availability of data). Meanwhile the best practical check of validity is to compare the results to expert expectations. To date, reviews of draft results have been made by the EP3 Safety study team and ATM specialists, and suggested changes have been incorporated in the modelling (as documented in Appendix II of (Ref. 1)). In future work, it would be desirable to



validate the results using wider groups of SESAR stakeholders during the SESAR Development Phase.

Recommendation # 8: *Review IRP initial predictions for SESAR with wider groups of experts in the subject, and ensure consistency between the IRP and the appropriate safety case. This may require a tailor-made training on the IRP (or equivalent accident-incident model) to ensure that each participant understands the approach to risk modelling.*

In addition, Safety Targets Achievement Roadmap (STAR) addressing the definition of the sequence of changes between the present and the planned future ATM system, enables to appraise whether the predicted safety improvements throughout the period are not outweighed by the extra traffic. Ultimately, the Roadmap will include safety monitoring targets, so that as OIs are introduced, it can be determined if expected safety impacts are realised, exceeded, or fall short. This will lead to a true risk management system based on operational feedback. This feedback could then be used to revisit initial predictions.

4.3.4 Achievability

Achievability involves checking that safety requirements derived from the model are reasonably practicable to achieve in practice. This is a specific type of convergent validity for IRP results in the form of safety requirements.

Verifying achievability is the final step in the methodology that is planned for apportioning target levels of safety to obtain safety requirements for ATM elements. Ideally, this should be through a process of negotiation between the IRP target apportionment and the safety case for the ATM element(s) (OI and/or set of OIs). The safety case should demonstrate how the apportioned target will be met. If it is too difficult to meet, or if it can easily be exceeded, the safety case should propose an alternative target that is achievable for the ATM element.

Achievability is unlikely to be simply defined. In general, greater attention to design of any ATM element can achieve better performance. Hence, a wide range of targets might be “achievable”, with greater or lesser complexity and effort. This effort ultimately translates into cost. Hence, there is a trade-off between safety performance and cost, which is usually encapsulated by replacing “achievable” with the phrase “reduced as far as reasonably practicable”.

Because it might take a long time to establish whether an ATM element can achieve a given target, achievability might be demonstrated by comparing actual in-service performance or actual design targets with the requirements. Unfortunately, these are rarely available. To the extent that in-service performance is available, it has been used in the IRP development. This means that no validation is possible without collecting further information. In future work, it would be desirable to collect this type of information. It could then be used as to progressively validate the IRP (by comparing with it) and update it (by being incorporated within it). This is a type of convergent validation.

Recommendation # 9: *Collection of in-service performance and actual design targets for ATM elements, in order to check achievability of safety requirements.*

4.3.5 Face Validity

Face validity refers to stakeholder acceptance of the model. This is an on-going process, indicated by whether stakeholders accept and make use of the model results.

There has already been shown interests from European ANSPs, which indicates their need for this type of model. IRP has been provided to the FAA under Action Plan 15, and the FAA is populating the model with US data. Although this work is at an early stage, this is a clear indication of external stakeholder acceptance of the model.

Further validation of this type is recommended:



Episode 3
D2.4.3-03 - Note On Risk Model Validation

Version : 1.01

- Evaluation of independent proposals for improving accident safety. The European Commercial Aviation Safety Team (ECAST) is a component of European Strategic Safety Initiative (ESSI). ECAST addresses large fixed wing aircraft operations, and aims to further enhance commercial aviation safety in Europe, and for European citizen worldwide. During ECAST Phase 1, eighteen safety subjects were identified as topics for further analysis in Phase 2. Once safety recommendations provided by ECAST groups are available, it would be desirable to evaluate these using IRP. If they were either consistent with conclusions from the IRP or different for a clear reason, this could be considered a validation of it.
- Experts review of the conclusions from the IRP. If the recommendations in (**Ref. 1**) for the initial SESAR risk picture were supported by wider groups of industry experts during the SESAR Development Phase, this could also be considered a validation of the IRP. In general, if the main components of IRP could be subjected to expert review, this would be a way of improving the model in the short-term, but would also lead to a very powerful form of validation in the longer term.

Recommendation # 10: *Review of the face validity of IRP and its conclusions with progressively wider groups of stakeholders during the SESAR Development Phase.*



5. CONCLUSIONS

This note has reviewed the status of validation of the IRP, using the word in its many different senses. It concludes:

- The model development process has been fully validated by ensuring adequacy in the following aspects:
 - Specification for the model
 - Resources for model development
 - Competence of the model developers
 - Technical quality of the model
 - Compliance with the specification
 - Verification of the model construction
- The usability of the model is being validated through on-going definition of:
 - Ranges of validity for model inputs
 - Sensitivities of results to model inputs and parameters
 - Confidence ranges in the results
- The model results have been validated to the extent that is practical at present through:
 - Calibration against ATM changes during the period 1990-2005
 - Empirical validation against independent estimates of ATM overall contribution
 - Convergent validity against available statistics on ATM-related incident rates
 - Face validity through acceptance of the model by a limited group of stakeholders

In future work, further validation would be desirable. It is recommended that this should include:

Recommendation # 1: Peer review for IRP with progressively wider groups of stakeholders. This should be facilitated by the production of two types of report: (i) analytical report with details of the model; and (ii) operational reports containing the results and recommendations.

Recommendation # 2: Explore the modelling of user inputs ranges in the IRP further development.

Recommendation # 3: Modelling in the IRP the complete set of ATM changes that have taken place since 2005 and update the calibration of the IRP model against historical trends.

Recommendation # 4: Progressive calibration against accident and incident experience as it accumulates.

Recommendation # 5: Calibration of risks for specific regions or units for which there is suitable accident or incident data.

Recommendation # 6: Further collection of alternative parameter estimates for use in defining confidence ranges.



Episode 3

D2.4.3-03 - Note On Risk Model Validation

Version : 1.01

Recommendation # 7: Documentation of convergent validity as accident and incident datasets are enlarged.

Recommendation # 8: Review IRP initial predictions for SESAR with wider groups of experts in the subject, and ensure consistency between the IRP and the appropriate safety case. This may require a tailor-made training on the IRP (or equivalent accident-incident model) to ensure that each participant understands the approach to risk modelling.

Recommendation # 9: Collection of in-service performance and actual design targets for ATM elements, in order to check achievability of safety requirements

Recommendation # 10: Review of the face validity of IRP and its conclusions with progressively wider groups of stakeholders during the SESAR Development Phase.



6. REFERENCES

- Ref 1. EC EP3, "D2.4.3-02 – SESAR Top-Down Systemic Risk Assessment", May 2009.
- Ref 2. EUROCONTROL, SSAP, The European Strategic Safety Action Plan, A Synopsis of the Implementation Master Plan, Issue 1, October 2004
- Ref 3. FAA/EUROCONTROL Cooperative R&D Action Plan 15: Safety: Terms of Reference (AP15-03-TOR). v. 1.3. June 2003.
- Ref 4. SESAR Consortium, "The ATM Target Concept", D3, DLM-0612-001-02-00a, September 2007.
- Ref 5. SESAR Consortium, "SESAR Master Plan", D5, D L M - 0 7 1 0 - 0 0 1 - 0 2, April 2008
- Ref 6. SESAR Consortium, "Work Programme for 2008-2013", D6, D L M - 0 7 1 0 - 0 0 2 - 0 2 - 0 0, April 2008
- Ref 7. SESAR JU, WP 16 - R&D Transversal Areas, Description of Work (DoW), v4.0, 17th December 2008
- Ref 8. EUROCONTROL, "Main Report for the 2005/2012 Integrated Risk Picture for Air Traffic Management in Europe", EEC Note 05/06, March 2006.
- Ref 9. EUROCONTROL, IRP/STAR 3-day Training Material, February 2009
- Ref 10. EUROCONTROL, CBT CD-ROM as a means of self-study and training on IRP and STAR, April 2009
- Ref 11. EUROCONTROL, "Methodology Report for the 2005/2012 Integrated Risk Picture for Air Traffic Management in Europe", September 2006.
- Ref 12. EUROCONTROL, "Methodology for a Safety Target Achievement Roadmap", EEC Note v1.0, May 2007.
- Ref 13. EC EP3, "D2.4.3-04 – Method for Units of Operations", to be produced, July 2009.
- Ref 14. EC EP3, Annex 1 – Description of Work, V3.0/Contract amendment N°3 Rev 2, June 2008
- Ref 15. SESAR Consortium, "The ATM Deployment Sequence", D4, DLM-0706-001-02-00 – January 2008
- Ref 16. EC EP3, IRP2008 Package (Excel Tool) supporting EC EP3, D2.4.3-02
- Ref 17. Safety Regulation Commission (2002), "Aircraft Accidents/Incidents and ATM Contribution", SRC Document 2, EUROCONTROL, v3, December 2002.
- Ref 18. Safety Regulation Commission (2007), "Annual Safety Report", SRC Document 43, EUROCONTROL, v1, December 2007.
- Ref 19. Ranter, H. (2006), "Airliner Accident Statistics 2005", Aviation Safety Network.
- Ref 20. EC EP3, "D2.4.3-01 - White Paper on the SESAR Safety Target", v1.2, approved, September 2008.



7. GLOSSARY OF TERMS

Term	Definition
ANS	air navigation services
ANSP	air navigation service provider
ATC	air traffic control
ATM	air traffic management
ATS	air traffic services
BBN	Bayesian Belief Network
CATS	Causal Model of Air Transport Safety
ConOps	concept of operations
CNS	communication, navigation and surveillance
ECAC	European Civil Aviation Conference
EEC	EUROCONTROL Experimental Centre
FAA	US Federal Aviation Administration
ICAO	International Civil Aviation Organization
IRP	Integrated Risk Picture
JU	Joint Undertaking
NRW	non-dimensional risk reduction worth
OI	operational improvement
SAM	safety assessment methodology
SESAR	Single European Sky ATM Research Programme
STAR	Safety Target Achievement Roadmap



8. ANNEX - UNCERTAINTIES IN THE IRP

8.1 UNCERTAINTY IN EVENT FREQUENCIES DUE TO DATA QUANTITY

A traditional (or frequentist) risk model assumes that events follow a Poisson distribution, which is a convenient simple distribution defined only in terms of the mean frequency. If a number of events n has been observed in an exposure t , the maximum likelihood estimate of the underlying frequency is:

$$\lambda = n/t$$

If no failures have occurred in the exposure, an estimate may be obtained from the 50%ile of the distribution as:

$$\lambda_{0.5} = \frac{1}{2t} \chi_{0.5,2}^2 = 0.7/t$$

where:

$$\chi_{0.5,2}^2 = 50\%ile \text{ of the chi-squared distribution with 2 degrees of freedom.}$$

This is equivalent to assuming “0.7 events” in the exposure to date.

The estimated standard deviation of this estimate is:

$$\sigma = n^{0.5}/t$$

If n is large, this may be used to obtain the confidence limits on the mean as $\lambda \pm 1.64\sigma$.

More accurate confidence limits for small n are obtained using the chi-squared distribution:

$$\text{Lower confidence limit } \lambda_{0.05} = \frac{1}{2t} \chi_{0.05,2n}^2$$

$$\text{Upper confidence limit } \lambda_{0.95} = \frac{1}{2t} \chi_{0.95,2n+2}^2$$

These are readily obtained from statistical functions in Excel.

8.2 UNCERTAINTY IN EVENT PROBABILITIES DUE TO DATA QUANTITY

A Bayesian risk model typically assumes that the uncertainty in the event probability follows a beta distribution, which is a convenient conjugate family of distributions, whose form is preserved during Bayesian updating. The beta distribution is defined by two dimensionless parameters:

- A first shape parameter α
- A second shape parameter β

The parameters α and β are equivalent to the number of failures and successes in the data.

The mean of the beta distribution, and hence the expected value of the probability is:

$$\mu = \frac{\alpha}{\alpha + \beta} = \frac{n}{t}$$

The standard deviation is:



$$\sigma = \sqrt{\frac{\mu(1-\mu)}{\alpha + \beta + 1}} = \sqrt{\frac{n(t-n)}{t^2(t+1)}}$$

The confidence limits on the beta distribution are readily obtained from statistical functions in Excel.

Lower confidence limit $p_{0.05} = \text{beta}(n, t-n+1)$

Upper confidence limit $p_{0.95} = \text{beta}(n+1, t-n)$



Episode 3
D2.4.3-03 - Note On Risk Model Validation

Version : 1.01

END OF DOCUMENT