



Performance Monitoring and Management Toolset Evaluation Report

D5.2

INTUIT

Grant:	699303
Call:	H2020-SESAR-2015-1
Topic:	Sesar-11-2015 ATM Performance
Consortium coordinator:	Nommon
Edition date:	7 May 2018
Edition:	01.00.00

Founding Members



Authoring & Approval

Authors of the document

Name/Beneficiary	Name/Beneficiary	Name/Beneficiary
Luca Piovano/UPM	Researcher	27/04/2018
Fran Luque Oostrom/UPM	Researcher	27/04/2018
Núria Alsina/ALG	Researcher	27/04/2018
David Toribio / Nommon	Researcher	27/04/2018
Rodrigo Marcos/Nommon	Researcher	27/04/2018

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
David Toribio/Nommon	Researcher	20/04/2018
Rodrigo Marcos/Nommon	Researcher	20/04/2018

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Ricardo Herranz/Nommon	Project Coordinator	27/04/2018

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
------------------	----------------	------

Document History

Edition	Date	Status	Author	Justification
00.00.01	20/04/2018	Draft	Luca Piovano	First version
00.01.00	27/04/2018	Approved for submission to SJU	Luca Piovano	Inclusion of additional evaluation exercises and approval by INTUIT Consortium
01.00.00	07/05/2018	Approved	Luca Piovano	Approved by the SJU

INTUIT

INTERACTIVE TOOLSET FOR UNDERSTANDING TRADE-OFFS IN ATM PERFORMANCE

This document is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 6993030 under European Union's Horizon 2020 research and innovation programme.



Abstract

The purpose of WP5 is to design and develop a performance monitoring and management toolset organised around the concept of an interactive dashboard equipped with a set of visual analytics tools. This document describes the evaluation of this dashboard. The evaluation has been conducted with a small set of participants with professional background and expertise in projects related to ATM performance, in order to assess the usefulness and applicability of the dashboard and derive recommendations for future improvement.

Table of Contents

Executive Summary	5
1 Introduction	6
1.1 Scope and objectives	6
1.2 List of Acronyms	6
1.3 References	7
2 The evaluation process	8
2.1 Introduction	8
2.2 Objectives	8
2.3 Methodology	9
2.3.1 Experiment configuration	9
2.3.2 Participants	10
2.3.3 Procedure	10
3 Analysis of results	24
3.1 Introduction	24
3.2 Analysis of the evaluation	24
3.2.1 User profiling.....	24
3.2.2 Exercise analysis	25
3.2.3 Evaluation analysis	26
3.2.4 Applicability to real operational environments	26
3.2.5 General remarks.....	27
4 Conclusions and future work	31

Executive Summary

This document describes the evaluation process followed in INTUIT to assess the performance monitoring and management toolset developed in WP5 and described in D5.1. This toolset has been conceived as a web-based visualisation dashboard aimed to help analysts and stakeholders explore the results provided by the ATM performance models developed in WP4 through a user-friendly interactive interface.

The evaluation of the aforementioned toolset has been conducted with a small set of participants with professional background and expertise in projects related to ATM performance, in order to assess the usefulness and applicability of the dashboard for performance analysis and decision making.

General qualitative opinions of the participants about the overall interface, the graphics presented, the ease of use and the performance of the dashboard have been gathered. This feedback has helped the developers to detect and understand usability issues related to the general layout, the interaction with the different components and the way information is presented to the user. The lessons learnt and the possible improvements of the toolset are discussed in the conclusion section.

1 Introduction

1.1 Scope and objectives

The INTUIT project aims to explore the potential of Visual Analytics (VA) and Machine Learning (ML) to improve the understanding of the trade-offs between ATM KPAs, identify cause-effect relationships between KPIs at different scales, and develop new decision support tools for ATM performance monitoring and management.

The present deliverable describes the evaluation process used to assess the prototype visualisation platform developed in INTUIT WP5. The general purpose of such visualisation environment is to provide a toolset based on the Visual Analytics paradigm for performance monitoring and management, organised around the idea of interactive dashboards. More precisely, the whole visualisation platform addressed the following specific objectives of WP5:

- to develop multi-objective optimisation engine to find Pareto optimal solution for a set of KPIs;
- to develop an interactive dashboard for multi-criteria and sensitivity analysis; and,
- to develop a prototype integrating the developed tools.

This document focuses on the demonstration and evaluation of the prototype by analysing some interactive session with potential users. The assessed prototype is described in D5.1 Performance Monitoring and Management Toolset. Such tool is a set of *ad-hoc* environments and suitable visual representations designed to allow analysts to interactively explore and analyse their modelling results. The data related to such results come from the case studies developed in WP4 (for further details, see D4.1 Performance Metrics and Predictive Models), namely:

- CS-1 Study of the effect of unit rates on en-route performance.
- CS-2 Identification of sources of en-route flight inefficiency.

Case study CS-3 Development of new multi-scale representations of ATM performance indicators was visually analysed by Fraunhofer IAIS group using their own toolset, V-Analytics. Due to some software incompatibility issues, their results were not integrated in the platform developed in WP5.

1.2 List of Acronyms

Acronym	Definition
ACC	Area Control Centre
ANSP	Air Navigation Server Provider
ATM	Air traffic Management

Acronym	Definition
CZ	Charging Zone
KPA	Key Performance Area
KPI	Key Performance Indicator
NM	Nautical Miles
PCP	Parallel Coordinates Plot
VA	Visual Analytics

Table 1 - Table with the acronyms used throughout this document

1.3 References

The present deliverable has been written in accordance with the following INTUIT documentation:

- Grant Agreement N. 699303 INTUIT – Annex 1 Description of the Action.
- INTUIT D1.1 Project Plan, v00.02.00, June 2016.
- INTUIT D1.2 Data Management Plan, v01.00.00, December 2016.
- INTUIT D2.1 Performance Data Inventory and Quality Assessment, v01.00.00, December 2016.
- INTUIT D2.2 Qualitative Analysis of Performance Drivers and Trade-offs, v01.00.00, November 2016.
- INTUIT D3.1 Visual Analytics Exploration of Performance Data, v01.00.00, October 2017.
- INTUIT D4.1 Performance Metrics and Predictive Models, v00.03.00, February 2018;
- INTUIT D5.1 Performance Monitoring and Management Toolset, v00.01.00, April 2018.

Furthermore, the following resources have been used as references:

- [1] Goldstein, E. B., & Brockmole, J. (2016). Sensation and perception. Cengage Learning.
- [2] Ware, C. (2012). Information visualization: perception for design. Elsevier.
- [3] Elmqvist, N., & Yi, J. S. (2015). Patterns for visualization evaluation. Information Visualization, 14(3), 250-269.
- [4] ITU-R. (2000). Subjective assessment of stereoscopic television pictures. International Telecommunication Union, Geneva.
- [5] RITUR BT (2002). Methodology for the subjective assessment of the quality of television pictures.
- [6] INSIGHT Consortium (2015). Analysis of Urban Location Patterns. Deliverable of the FP7 European Project INSIGHT.
- [7] Galloso, I. (2015). User Experience with Immersive and Interactive Media. PhD Thesis.
- [8] Likert, R. (1932). A technique for the measurement of attitudes. Archives of psychology.

2 The evaluation process

2.1 Introduction

Human beings are provided with a perception mechanism that acts like a communication bridge between the external world and the self. Without loss of generality, the interpretation of the information captured by the different sense organs can be considered as a two-stage process: first, the early sensory processing level focuses on extracting relevant features from the incoming multimodal sensory information; then the high-level cognitive processing level aims to consciously interpret and judge such refined information [1].

Visual analytics approaches aim to enhance these perceptive mechanisms, such that the mutual links between the low- and high-level functionalities foster the cognitive acquisition of new information. In this sense, the design of the visual interfaces for data exploration and understanding is particularly critical since it must support this perceptive flow in a coherent and effective manner. The literature specialised in human factors has produced a lot of evidence about the different roles played by a variety of technical and contextual influences that affect positively and negatively these processing functions [2]. At the same time, the evaluation of visualisation tools seems more challenging because the high-level tasks generally supported by visual analytics (e.g., make a complex decision, extract insights, understand phenomena) are difficult to isolate, characterise, and quantify [3].

In INTUIT, the visualisation platform provides several interactive dashboards for the exploration and understanding of ATM performance-related questions. Assessing the accomplishment of this goal is the main subject of this document. In it, we describe the methodology used to evaluate the effectiveness of the visual environment. The proposed methodology is based on an empirical evaluation where a group of participants is asked to answer a set of questions involving the interaction with a dashboard and simulating a possible decision-making process. The main features to analyse comprise usability and tasks completion. The users' evaluation and comments as well as the analysis of their interaction with the different graphical components allow us to draw conclusions about possible improvements. This is particularly important in view of a potential future version of the visualisation tool adapted for its application in an operational environment.

2.2 Objectives

The goal of the evaluation process is twofold:

- to study to what extent the visualisation tools and techniques implemented in WP5 are effective in supporting a possible decision-making problem;
- to understand how the user perceives the overall quality of the dashboard and its elements.

2.3 Methodology

2.3.1 Experiment configuration

Due to time and resource limitations, the evaluation experiment has been designed to achieve a trade-off between three different factors, namely:

- the complexity of the task to be performed (in terms of number of widgets to interact with and the number of steps required to come with a solution);
- the meaningfulness of the task (i.e., a possible outline of a real problem to tackle in a real operative environment); and,
- the duration of each session of experiments (i.e., to keep it as short as possible).

As the entire process should not take more than one hour, the required balance put some constraints on the experiment design process. First of all, we decided to analyse only a single dashboard, namely the one corresponding to the CS-1 case study. This is at the same time the most complex and complete of the two dashboards developed in WP5¹, and therefore it appears to be the natural choice for the sake of the assessment process. Based on this choice, the final configuration of the evaluation session is based on the analysis of some optimisation solutions with respect to an initial scenario, by exploring their general effects on different indicators and the main benefits and drawbacks for the stakeholders involved. This task has been divided into a set of eight questions guiding the user through the process of the exploration of the policy effects. There are essentially two classes of questions: the first one is expected to have a numerical answer; the second one allows a more descriptive reply. The exercise part is estimated to last between 15 to 20 minutes. The same evaluation exercise is given to each participant involved in the assessment process in order to have a common basis to analyse the evaluation results.

With the aim of providing an experience as natural and fluid as possible, the order of the questions follows two general criteria. First, the tasks to be completed must have increasing difficulty and complexity to enable a progressive practical enhancement of the individual's skills with the dashboard interface. Second, the sequence of steps to be followed must match the logical order of a typical decision-making workflow.

To get the participants more familiar with the dashboard, prior to the exercise the user is asked to watch a video tutorial where the main functionalities of the dashboard are presented. The first goal of the tutorial is to have a uniform reference across the different users and interviewers involved in the evaluation process. This way, it is guaranteed that each participant is provided with exactly the same information at the beginning of each session. This is particularly important when the interviewees have no prior experience with the tool and different interviewers are involved in the process. After the video session, the user can try on its own the functionalities of the dashboard. The interviewers can also answer the questions a participant may have about the visualisation

¹ The CS-2 dashboard presents a subset of the functionalities and visualisation tools included in CS-1 visual environment.

environment. A total of four people have been involved into the evaluation process as interviewers (2 people from Nommon and 2 from UPM-CeDInt).

Section 2.3.3 describes the overall procedure used for the evaluation purpose. This will include the following steps: an initial text with a general description of the dashboard to provide a global context in terms of data and objectives (Section 2.3.3.1), the tutorial session to get familiar with the dashboard (Section 2.3.3.2), the exercise itself (Section 2.3.3.4), and the final questionnaire the user is asked to fill in.

2.3.2 Participants

All the participants involved in the evaluation process were chosen to have some experience about the topic dealt with the dashboard to assess (i.e., optimisation problems in the airspace domain). A variety of profiles was sought, in order to collect different points of view.

The total number of volunteers recruited for the experiment is 6. Their expertise fields are described in Section 3.2.1.

No personal data were collected, but those concerning their professional profiles.

People directly involved in the project were not eligible for the assessment. On the other hand, some of them participated as interviewers. People working for some consortium partner were allowed for the experiment if and only if they were never exposed to the INTUIT platform.

2.3.3 Procedure

The self-developed procedure used for the evaluation was agreed between Nommon and UPM-CeDInt groups. It has been inspired by the ITU-R Recommendations BT.1438 and BT.500 ([4],[5]), and by previous research and experience in Human Computer Interaction by the UPM-CeDInt group ([6],[7]).

Prior to the experiment, the participants were provided with an informed consent to sign and notified that no personal data (apart from their job position and the years of experience) would be collected for the purpose of this study. This also implies that the collected answers and feedback are anonymised. All participants were tested individually through a dedicated Skype call: the interviewee shared his screen to allow the interviewers to follow the exercise development. All the experiments were conducted avoiding any interruption.

Before the evaluation process started, the participant was invited to sit down comfortably. Then the interviewer quickly described the purpose of the evaluation session and what would have happen in the following steps. After the beginning of the experiments, the interviewer took note of any interesting issue occurred (e.g., about interaction, issues within the platform, user's reaction), especially during the exercise phase.

The experiment was conducted in 4 separate phases. The details for each phase are explained below.

2.3.3.1 Introduction

The selected user is provided with a small, introductory text describing the scope and the objectives of the dashboard. The text is the following:

“This dashboard provides a tool to assess the performance effects of policies affecting three Key Performance Areas (KPA): cost-efficiency (in terms of *Navigation costs*), capacity (in terms of *Delay*) and efficiency (in terms of *Route extension*). The data used for this purpose come from a route choice predictor: it is used to calculate aggregated efficiency metrics for a given route as a function of the unit rates set in the different charging zones, which in turn determines the charges paid to fly a route. The dashboard enables the evaluation of the trade-off between flight efficiency, cost efficiency and capacity by means of different interactive visualisations and the assessment of the effect of unit rates and route choices on ATM performance.”

The words written in italic refer to the names as shown in the dashboard.

The time allocated to finalise this part is about 2 minutes.

2.3.3.2 Tutorial session and user exploration

In this part, the user is invited to watch the video tutorial that the developer team recorded for the purpose of this assessment task.

After this video session, the user is asked to try the dashboard on his/her own in order to get familiar with it. During this period, the interviewers can answer the possible questions arising from the users.

The time slot for this part is about 20 minutes: more or less 15 minutes are devoted to the video tutorial session and the remaining part for the trial and question phase. In some cases, this last part lasted more than 10 minutes because the volunteers asked for more explanations about the representations and how to interact with them and/or the underlying data represented.

2.3.3.3 Exercise

The third part is devoted to practically solve some specific tasks by directly interacting with the dashboard. To this end, the user is provided with a decision-making problem to solve, split in a set of questions to answer. Each question either describes a task to perform or requests the user to provide his/her own opinion about the situation depicted in the dashboard itself. Although the problem could be thought as a simplification of a possible, real decision-making scenario, the sequence of questions is intended to mimic the main steps of a typical decision-making routine, by interchanging explorative/inquiry steps with thinking/speculative moments. The questions prepared for this part are provided below (words written in italic refer to labels present in the dashboard):

- Go to the *Optimisation Window* and set the *Criteria Filters* as follows:
 - *Route extension*: between *Low* and *High*. Uncheck *baseline*.
 - *Navigation costs*: set to *None*. Uncheck *baseline*.
 - *Delay*: set to *Medium*. Uncheck *baseline*.

Note that these values will be translated to the solutions as numeric values:

- *None* = 0
- *Low* = 1
- *Medium* = 2
- *High* = 3

Baseline scenario is identified by -1/-1/-1 values

- Among the filtered options, which solution decreases the most the *route_extension* indicator?

- With respect to the baseline scenario, what are the effects of this solution over the other indicators?

- Consider the filtering criteria applied in step 1. If you wanted to find a good trade-off between *Route extension* and *Regulations*, which solution would you pick?

- Explain why you selected this option:

- Which chart(s) did you look at to make your choice? Why?

- Configure the *Criteria Prioritisation* to show the solution picked in exercise 4 and then go to the *Analysis Window*. With respect to the baseline scenario, what consequences has that option in terms of *Unit rates* for each charging zone?

- Analyse the equity of the selected solution for the affected *Airlines*: Which indicator is the most affected? Regarding this indicator, is there any airline benefiting the most from the adoption of this policy option? Which plot/data gives you that information?

The time estimated to complete the exercise is between 15 and 20 minutes.

2.3.3.4 Questionnaire

The last part of the user exercise concerns the assessment of the visualisation dashboard. To this end, a questionnaire was provided to the interviewees with the aim to collect their quantitative and qualitative feedback.

Quantitative and qualitative evaluation:

The questionnaire has been divided into three different sections. In the first one, the user is asked to score the interactive environment from the point of view of its presentation, layout and usability. The rating scale chosen for this assessment is a five-level Likert-type scale (see [8]), where the two end-points, 1 and 5, represent the worst and best grades, respectively. Beside this numerical judgement, the user is asked to provide some free, textual comment, if any. The idea is to have a more complete picture of the users' impressions in order to find possible shortcomings to improve and refine. The points to score are listed below, in the same format as they have been provided to the interviewee.

- Give a score between 1 (worst) and 5 (best) to the following features and provide some additional comment, if any:

- Overall look and feel (your general judgement about the visual impact of the visual environment).

Score:

Comments:

- Graphical presentation (Is the layout ordered, clean and pleasant? Are the colours clearly distinguishable?)

Score:

Comments:

- Text readability (Are the labels easy to read? Do you think the font and size are adequate?)

Score:

Comments:

- Ease of use (How simple was it to accomplish the exercise task?)

Score:

Comments:

- Content coherence (Is the information presented across the widgets and charts coherent? Do you spot some incongruence?)

Score:

Comments:

- Meaningful representations (Are the graphical tools good enough to present the data? Was it easy to draw the conclusions required for the exercise?)

Score:

Comments:

- Stability (Did you experience any problems – server disconnection, slowness in response, bugs – while using the tool?)

Score:

Comments:

- Overall user experience.

Score:

Comments:

Operational environment evaluation:

The second part of the evaluation phase is aimed to identify the potential uses of the prototype in a real, operational environment. The tool developed within INTUIT WP5 has been built for research purposes to support a specific subset of problems. Although it has been implemented having in mind the idea of possible extensions adapted to the needs and requirements of a real environment, the tool is still not ready for an industrial environment. To be concretely applied in a proper industrial environment, some issues have to be tackled before, such as: an overall generalisation of both the content presentation and the range of problems to be studied; the enrichment of the functionalities to provide more flexibility and, at the same time, expressiveness strength; and avoiding specific concerns like the system response speed, layout imperfections, and usability issues.

In this context, the answers provided in this section are expected to be general hints towards this objective. The experience of the interviewed experts will shed light on the possible future direction of development.

The questions asked in this section are listed below.

- Do you see any use of the tool in your organisation?

- What do you see as the main application of the tool?

- Where do you see a potential use of this kind of tool?

- How would you adapt the tool for any other use?

Additional comments:

Finally, the last part deals with a section for general comments where the user is invited to give more information about the topics not covered in the two previous sections.

2.3.3.5 Exercise solutions

For the sake of completeness and to better understand the kind of answers expected, in the following the solutions to the exercise described in Section 2.3.3.3 are provided.

1. Go to the *Optimisation Window* and set the *Criteria Filters* as follows:
 - a. *Route extension*: between *Low* and *High*. Uncheck *baseline*.
 - b. *Navigation costs*: set to *None*. Uncheck *baseline*.
 - c. *Delay*: set to *Medium*. Uncheck *baseline*.

Note that these values will be translated to the solutions as numeric values:

- *None* = 0
- *Low* = 1
- *Medium* = 2
- *High* = 3

Baseline scenario is identified by -1/-1/-1 values

The solution is provided in Figure 1 below.

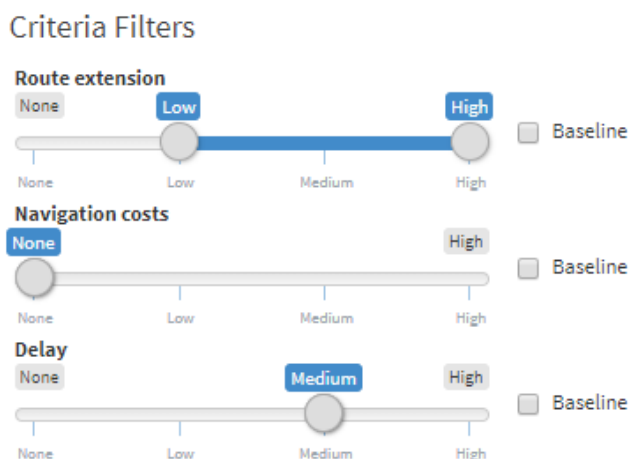


Figure 1 - Setting the criteria filters as asked in the formulation of the first question of the user's exercise

2. Among the filtered options, which solution decreases the most the *route_extension* indicator?

The answer is the solution highlighted in blue (3/0/2). The filtering provided in the first question leaves at the analyst's disposal only three solutions plus the baseline and the selected ones (always present by default). A possible manner to find the correct solution is by hovering the cursor over the different lines in the PCP chart until the one with the lowest *route_extension* value is highlighted. Another possible approach would be ordering the data table by the *route_extension* dimension and comparing its values against each other to find the requested answer. In this case, this value comparison is feasible since the number of values to compare is small. In general, it is better to rely on the PCP (with some extra filtering, if required, to reduce some visual cues). A screenshot representing the dashboard configuration to answer this question is given in Figure 2.

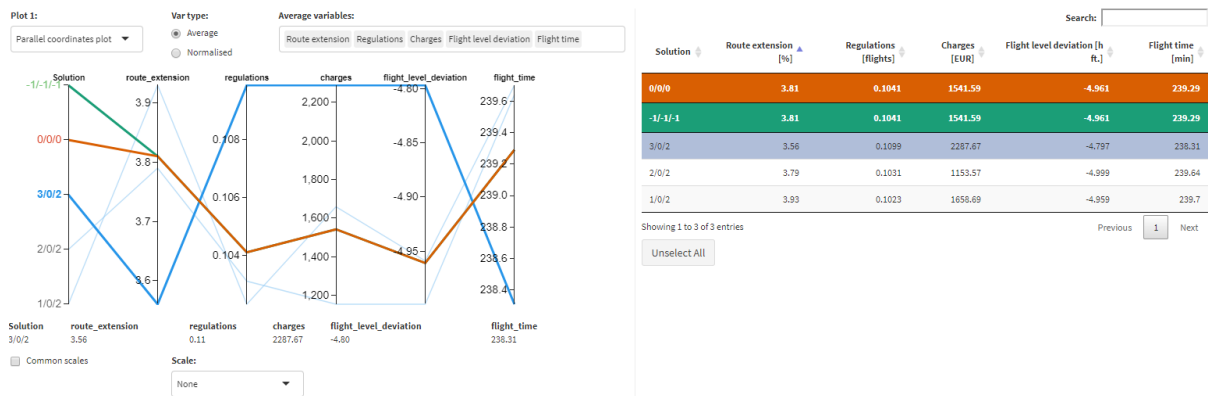


Figure 2 - The situation of the dashboard to answer the second question of the user's exercise.

3. With respect to the baseline scenario, what are the effects of this solution over the other indicators?

The overall effects of the solution minimising the route extension indicator can be observed in the leftmost chart of Figure 2 or in Figure 3: the average number of regulations, the average charges to pay and the average flight level deviation will exceed the corresponding values of the baseline, meaning that these solutions will produce a worse situation with respect to these dimensions. On the other hand, the flight time indicator considerably improves with respect to the baseline scenario.

Figure 2 shows the PCP configuration when each column has its own, original scale (that is, representing the original values). By playing a little bit with the PCP selectors, the analyst may represent the same set of solutions by taking as a reference the baseline itself, as shown in Figure 3. In this case, the numbers on the vertical axes deal with normalised values and a common scale is set. The baseline scenario is represented as a straight line. Even if the representation has changed, the general properties of the solutions therein are kept unchanged.

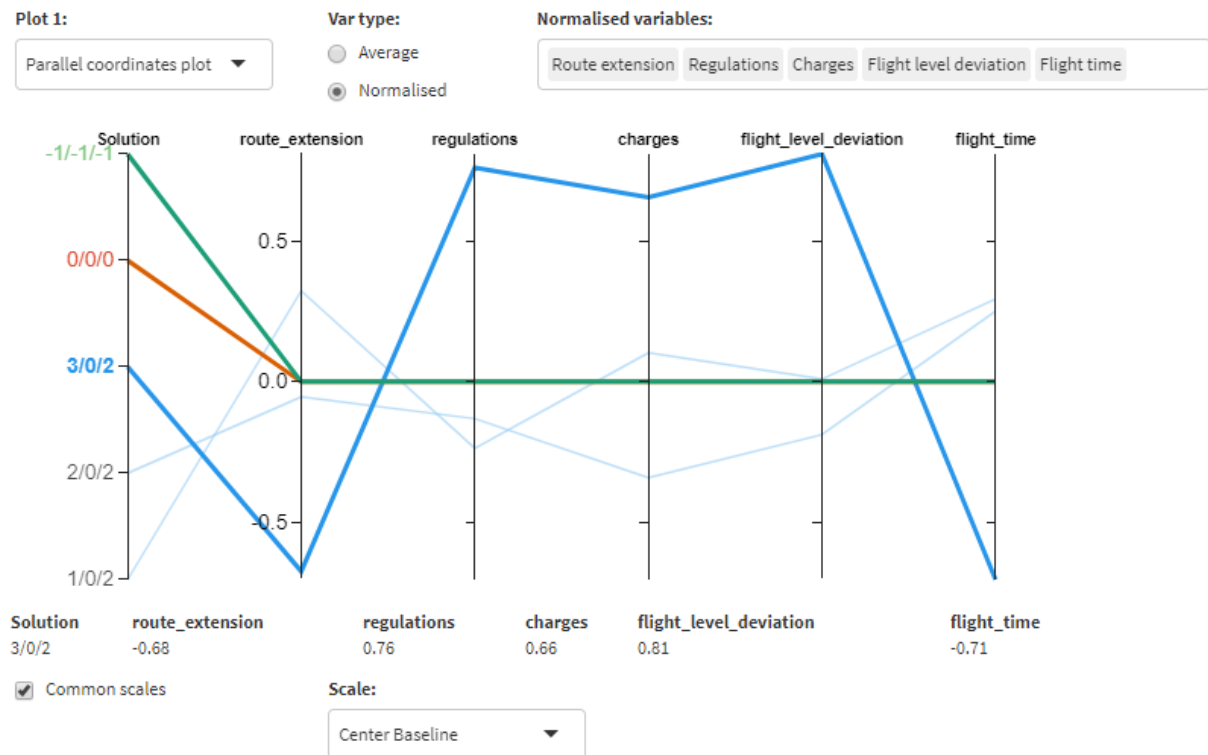


Figure 3 - Playing with the PCP to derive similar conclusions for the second question.

4. Consider the filtering criteria applied in step 1. If you wanted to find a good trade-off between *Route extension* and *Regulations*, which solution would you pick?, and
5. Explain why you selected this option

The right solution to this question is identified with the ordered triplet 2/0/2 (see Figure 4). To determine it, the idea is to see where there is a balance in the PCP columns corresponding to the indicators mentioned in the question. Among the solutions at the analyst's disposal, the one identified as 3/0/2 has the minimum value for the route extension indicator but the highest value for the regulation dimension. The solution 1/0/2 shows an opposite but less extreme trend: the route extension is worsening the baseline scenario while the regulation indicator is improving it (by the way, the best improvement of all the solutions under analysis). Solution 2/0/2 is improving the starting situation along both dimensions, and for this reason it represents the best option for this specific case.

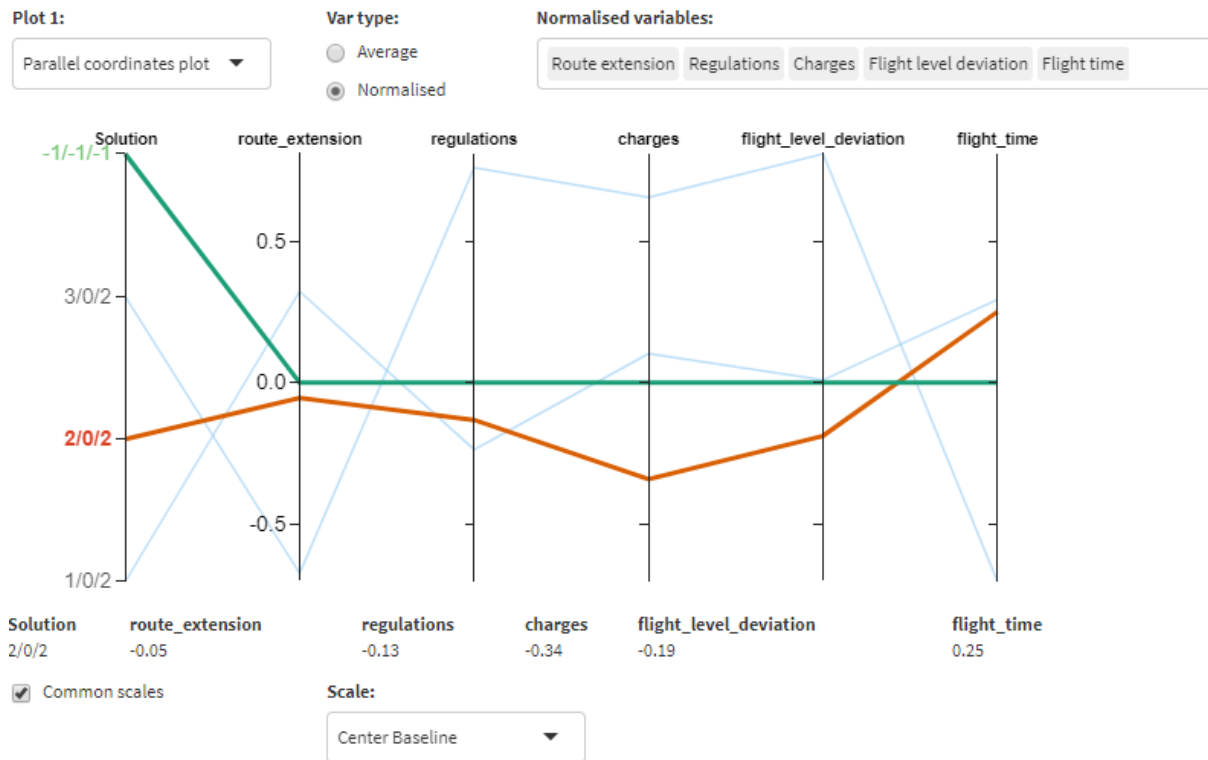


Figure 4 - The PCP used to answer to the fourth and fifth questions of the exercise. The orange line corresponds to the solution of this question.

6. Which chart(s) did you look at to make your choice? Why?

The previous question could be answered by looking at the chart in Figure 4, as explained before. An alternative could be looking at the scatterplots depicted in Figure 5 and Figure 6. In the first case, the Pareto optimal solution option is enabled in the chart selector: the best option for the exercise scenario is the one in the middle of the chart. In the second case, the solutions are scaled with respect to the baseline. With this representation, the solution laying in the South-West quarter is the best one since it improves the performances of the baseline in both dimensions at the same time.



Figure 5 - The scatterplot configuration that can be used to answer the question proposed in the fourth point of the exercise: the selected solution is represented surrounded by an orange square. This represents an alternative view to the one depicted in Figure 4



Figure 6 - Another possible way to answer questions number 4, 5, and 6: the solutions are presented scaled with respect to the baseline (as in the PCP in Figure 4). The point in orange is improving the two indicators in both dimensions and therefore it is drawn at South-West w.r.t. to baseline (in green).

7. Configure the *Criteria Prioritisation* to show the solution picked in exercise 4 and then go to the *Analysis Window*. With respect to the baseline scenario, what consequences has that option in terms of *Unit rates* for each charging zone?

The map in Figure 7 displays the effects of the selected solutions by representing the differences with respect to the baseline scenario. In particular, the areas coloured in green shades will have an increase in their unit rates: among them, Morocco is the one with the greatest increment. On the other hand, the regions represented in brown will show a decrease in their unit rates: France, the United Kingdom and Spain are the charging zones whose decrement is most evident. The charging zone of Ireland does show any relevant change with respect to the baseline.

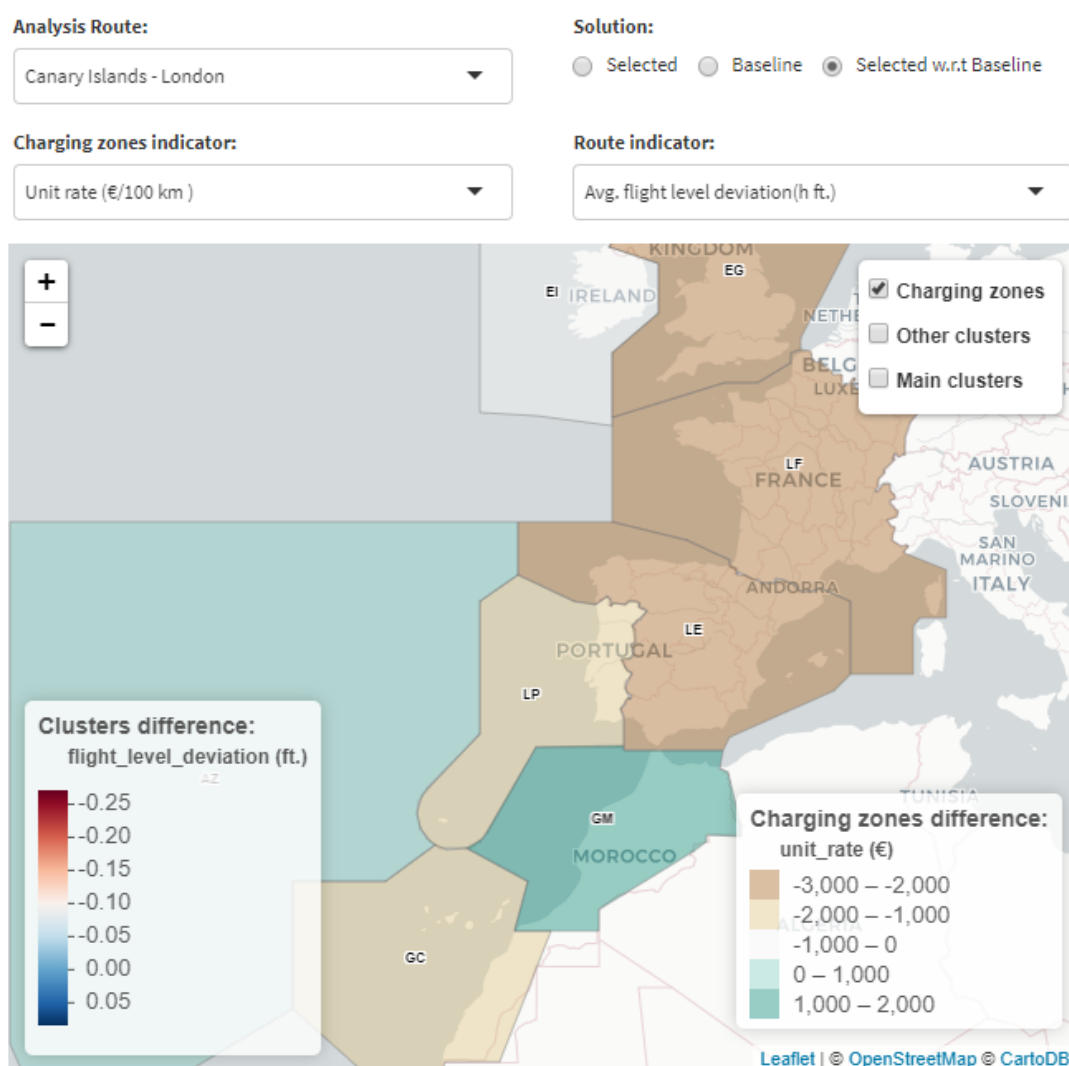


Figure 7 - The map where to compare the effects of the policy 2/0/2 w.r.t. baseline scenario.

8. Analyse the equity of the selected solution for the affected Airlines: Which indicator is the most affected? Regarding this indicator, is there any airline benefiting the most from the adoption of this policy option? Which plot/data gives you that information?

Figure 8 shows the overall situation for all the airlines included in the case study. The scatterplot represents the normalisation values because even the smaller effects are more visible. From the chart below, it seems that every airline is experiencing a decrease in the charges to pay, which is the indicator with the most evident and positive impact. Even if there are negative values next to each point, the correct interpretation is that they have a decreasing effect with respect to the baseline. All the other indicators experience some variations, but the overall magnitudes are not as big as the one shown for the charge dimension. From visual inspection, this is evident by looking at the size of the points for each dimension.

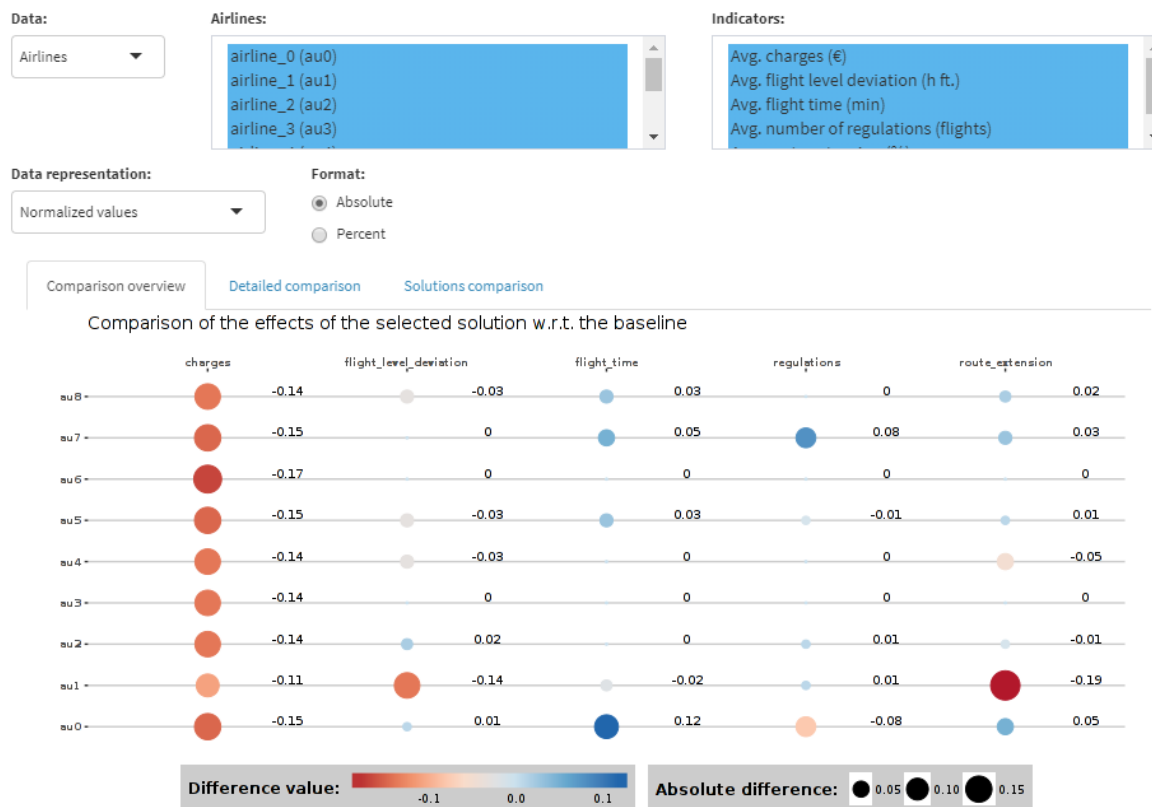


Figure 8 - The chart that could be used to analyse the effects of the selected options over the different airlines to answer to the question number 8.

Concerning the airline benefiting the most from the adoption of the policy 2/0/2, the correct answer is *au6*. The shade of red associated to the corresponding point is the darkest one. Also reading the numeric value close to this point confirms this suggestion: the -0.17 value is the smallest among all the values presented in the chart.

An alternative point of view is displayed in Figure 9, where a bar chart is used to infer the answers.

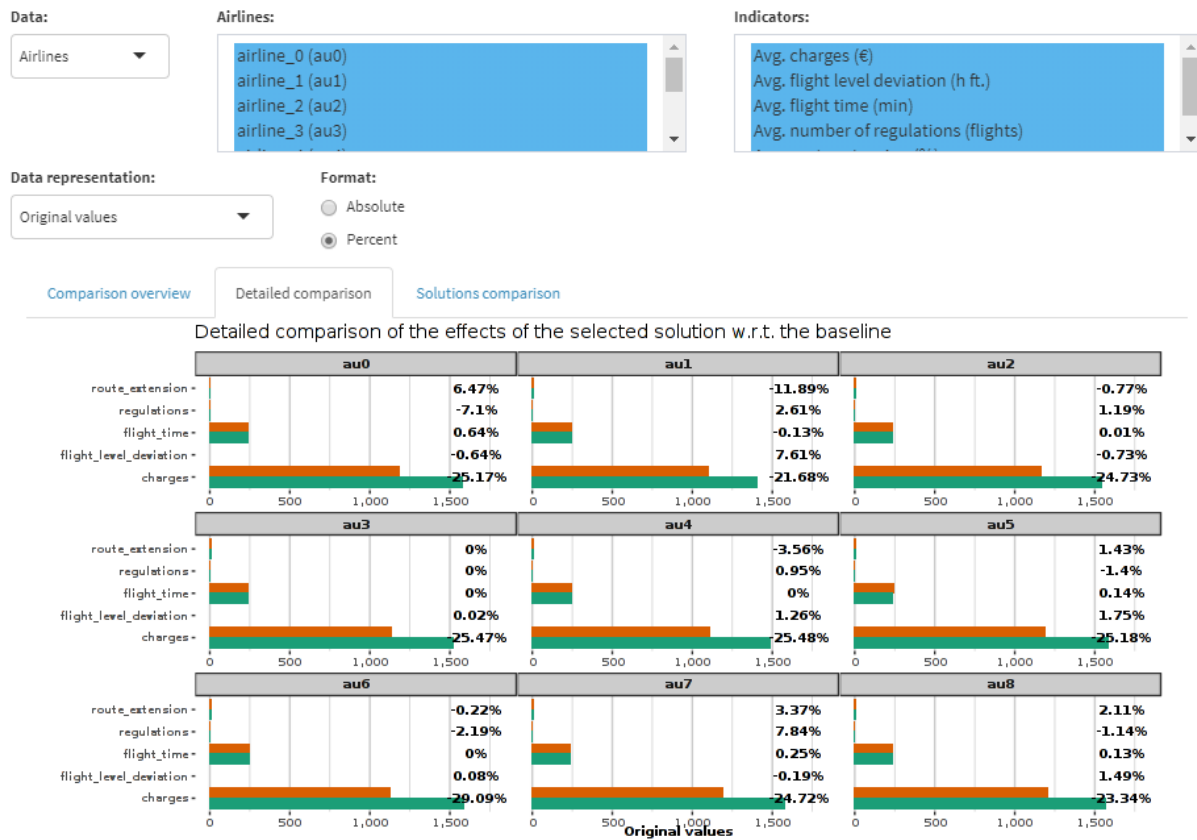


Figure 9 - Alternative chart to answer question number 8: in this case, the chart used is a bar chart representing the effects of both the selected solution and the baseline scenario. The numeric labels close to the bars represent the difference between the two scenarios as percentages.

3 Analysis of results

3.1 Introduction

In this chapter, a general discussion of the results of the evaluation exercise and the answers provided to the assessment questionnaire is provided. In particular, the focus here is to analyse how the visualisation tool has been used and how it has been perceived by the participants in terms of usability and look-and-feel.

The discussion is organised in five sections, each of them dealing with a specific aspect of the evaluation procedure.

3.2 Analysis of the evaluation

3.2.1 User profiling

For the purpose of the platform evaluation, no personal details of the volunteers involved in it have been collected, but their job profiles. The people participating in the evaluation session belong to one of the following profiles:

- A young consultant with a few years of experience in topics like airport operations.
- A consultant manager with a long experience in aerospace and digital innovation for the aviation sector.
- An operations manager with long experience in projects related with the SESAR program and security for ATM.
- A young software developer with experience in dashboard development in the context of ATM.
- Two senior researchers both with experience in SESAR projects related to trajectory optimisation and performance.

Therefore, all the participants have a technical profile and expertise in different domains related to aviation. This makes them particularly suitable for the technical evaluation of the visualisation environment. On the other hand, although some of them have participated in projects where one of the outcomes was a visualisation tool, in general their daily routines do not involve the use of such interactive applications. Despite this, their qualitative assessments concerning typical user-interface features is valuable as well, since it allows us to estimate several usability concepts and how steep it could be the learning curve before using the tool properly.

Regarding the platform, it is worth saying that the users tested the tool using different operating systems (Windows, MacOS and Linux/Ubuntu) and Internet browsers (Chrome, Firefox and Safari).

3.2.2 Exercise analysis

In general, the exercises were successfully completed by the volunteers in the foreseen time range, which is between 15 and 20 minutes. One of the users, however, could not complete the exercise in time, partially because some error prevented the login during several minutes.

Participants did not have any remarkable difficulty to understand the proposed questions (see Section 2.3.3.3). Therefore, we can conclude that the formulation of the different exercise tasks does not represent any hurdle for their thorough interpretation and correct execution. However, four of the volunteers misunderstood the difference between the criteria prioritisation selector and the criteria filters slider (this usability concern has been analysed in detail in section 3.2.5). In one case, this issue drove the user to analyse the whole set of optimisation solutions instead of the filtered set and in consequence the answers to the questionnaire did not match the expected ones.

In other case, a participant provided the wrong answer to question number 4, which in turn led him to provide likewise wrong responses to the two last questions. The error arose from an initial misinterpretation of the solution representations provided in the PCP of Figure 2: the participant indeed chose the default selected solution (0/0/0) thinking this was the best possible trade-off between the *Regulations* and *Route extension* indicators, as requested. Despite indicating this answer, the interviewer in charge of following the volunteer's exercise execution noticed that he/she was not completely convinced of his/her choice, especially when comparing the information provided by the PCP with the scatterplot and the Pareto optimal solution option. When asked about the choice taken, the participant provided a thorough and logical reply, demonstrating this way that he/she could really understand what the charts were representing, but was not able to perceive where the fault was and, therefore, to amend his/her own answer. A first conclusion about this fact suggests that it could be useful to add some functionality that automatically provides an explicit representation of the best solutions for a given set of conditions. This way, the analyst's judgement would be guided towards the correct identification of the best option at his/her disposal. This addition would be useful for both the PCP and the scatterplot representations.

Another user misunderstood question 7, thinking that the analysis had to be done regarding the impact on the average charges per cluster instead of the unit rate change per charging zone. Moreover, the user did not understand the meaning of "Main clusters", thinking that it referred to a single route instead of a classification meaning the most flown clusters.

Some indicators were misleading for some users. For instance, the indicator of "average flight deviation of flight level" is misleading as it is negative by definition. An improvement would be to present it in absolute value. Another issue pointed by some user was the definition of "route extension", which should be explained in some tooltip.

A general repeated question asked about what did "solution" and "baseline" stand for. This problem would have required a more comprehensive explanation of the methodology used to generate the data, which was obviated because of time concerns.

3.2.3 Evaluation analysis

Question	User1	User2	User3	User4	User5	User6	Average score
Overall look and feel	5	4	5	4	5	5	4.67
Graphical presentation	4	4	4	4	4	5	4.17
Text readability	4	4	4	3	4	4	3.83
Ease of use	3	5	3	4	5	5	4.17
Content coherence	5	4	4	5	5	4	4.50
Meaningful representations	4	4	4	4	5	5	4.33
Stability	5	5	5	4	3	5	4.50
Overall user experience	5	4	4	4	5	5	4.50

Table 2 - Details of the scores obtained in the evaluation process. The scale ranged from 1 to 5, being 5 the best scale rate. The evaluated points concern those features involving both look-and-feel and usability topics.

The readability of the text presented by the tool was rated with the lowest score of the evaluation. A general comment of the participants was that the font size employed by some components (e.g., the Criteria filters control and the Comparison overview chart) was too small to comfortably read the text. Since the dashboard is presenting an extensive amount of visual and textual information together in a limited space, the font size has been carefully chosen to prevent undesired effects like overlapping. Yet, font size employed by the components is currently fixed; a dynamic allocation of the font size for each component according to the available blank space should provide improved usability.

Also, it is interesting to note that another main concern about the dashboard is related to its ease of use, given that it has been rated twice with a score lower than 4. To interpret this assessment, two interrelated conclusions can be drawn. First of all, the volunteers directly played around with the proposed dashboard for about thirty minutes (ten minutes of free trial plus about fifteen minutes to complete the exercise), which is a small amount of time to get familiar with a new tool. An extended use would certainly improve the analyst's ability to interact with the different charts and widgets, such that he/she could entirely exploit the potential and capabilities of the dashboard. In addition, the readability achieved a lower-than-average score, maybe due to the use of laptop computers, which was not recommended.

3.2.4 Applicability to real operational environments

Concerning the potential applications of the research dashboard developed in WP5, some of the answers collected so far highlighted that it could be particularly valuable in optimisation problems, especially from the point of view of some specific stakeholder. For instance, several volunteers answered that airline route optimisation would be a possible target of the dashboard. More specifically, most of the participants identified the problem of understanding trade-offs between several indicators as one of the most prominent application fields for its deployment in a real operational environment. Another suggested application would be to compare user-defined routes.

The introduction of this type of visual analytics environment in the daily working routine is considered favourable and advantageous. Some users highlighted that the tool could be used for some air traffic- or airline-related research projects, especially to analyse the influence of the solutions on ANSP revenues. The visualisation tool would bring much more value if it allowed different geographical representations: for instance, not only charging zones or ANSPs, but also FABs. The volunteers were not introduced to the dashboard developed for CS-2, so they are not aware that such feature has already been introduced there, where the user can choose between charging zones and ACC representations. Another comment suggested extending the geographical representation to airspace outside the European skies.

Concerning possible extensions of the proposed visualisation dashboard, a participant suggested the idea of studying the effects of changes in ANSP unit rates when air traffic problems arise.

At a first glance, the dashboard raised positive comments about its potential to be extended to a real, operational environment.

3.2.5 General remarks

In the following, we present some of the remarks collected during the experiments. These notes have been made either by voice during the free trial (after watching the video tutorial) or written down in some of the text boxes present in the exercise and evaluation documents. For each of them, some possible actions are detailed to overcome the issue.

Usability concerns

- *The Criteria Prioritisation bar in the lateral bar could be understood as a filter instead of a way of selecting the solution:*
Both the Criteria Prioritisation and Criteria Filters sections are composed by a set of three sliders and a set of buttons/check-boxes to enable/disable the baseline solutions. The very similar layout could generate confusion when choosing the most suitable tool to use. A possible solution could be to differentiate the Criteria Prioritisation layout by allowing the user to choose the indicator weights through drop-down menus. Indeed, sliders are more inherently related to filtering while lists suggest picking some element from a choice list.
- *The application has in general too many filters:*
The application offers a series of filters and options in each of the graphs, which can be seen by a novel user as overwhelming. One option to reduce complexity would be to reduce the options for customisation of graphs. For instance, the “common scales” check could be hidden or disabled when using “average values” in the PCP.
- *Filtering on PCP is useful but not intuitive since there is not any explicit hint or advice about it:*
The procedure for filtering on PCP was described in the video tutorial and an example about how to do it was shown too. At the same time, it is true that this capability is somehow hidden, and, especially for novice users, this feature could be hard to find and use. The easier way to overcome this problem is to put an explanatory section explaining the main functionalities of each chart and how to make the best of it.
- *In some cases, the meaning of the numerical format used to represent the solutions is unclear:*
Although there is a legend in the lateral bar describing the correspondence between the solution and the criteria weights, some users find it difficult to correctly identify the relationship between the criteria and such numbers. Again, an explanatory section would be

helpful to solve this issue (e.g., by explicitly stating the meaning of each component of the criteria triplet). Including tooltip information when hovering over a specific solution would also be of help.

- *More information about how the different parameters are correlated and the nature of the data would be appreciated for a better understanding:*

In the dashboard the descriptive information is limited to an introductory text that the user can read before using the application. During the trial, special care was taken to describe to the volunteers the meaning of the different parameters and options. Nevertheless, and in order to present the tool to a broader audience, this information will be included in a reference manual that can be easily accessed during the analysis process. On the other hand, the focus of this dashboard is to present the results of a specific model and explore the solutions produced by it. In this sense, the potential user is an analyst already possessing this knowledge behind the scene. Nevertheless, this user's comment is a strong reminder to include such information section that could be useful in a more general and/or real operational context.

Charts issues

- *To evaluate trade-offs, the Pareto graphs may not be straightforward for a clear understanding:*

One participant found it difficult to use the scatterplot (specifically with the Pareto optimisation solution option) to derive insights about trade-offs. However, despite this difficulty, the volunteer was able to provide the correct solution to the question proposed. This suggests that, with the proper training, the analyst is able to discover the relationships depicted in the chart, even if the graphical element falls outside his/her familiarity boundaries. Nonetheless, the introduction of some extra help (e.g., a help tutorial, explicit and unequivocal chart labels) would be helpful to facilitate a thorough understanding of the graphical elements.

- *The diverging colour scale used in the map to encode the comparative results for the routes is somehow misleading when interpreting the results:*

As depicted in Figure 7, the diverging scale is a red-white-blue scheme, where red is associated with the lowest values and blue with the highest ones. In the case under analysis, the values in red are negative, but at the same time they represent an improvement over the baseline scenario (i.e., the selected solution has lower unit rates than in the reference case). In this context, confusion may arise since red is usually associated with a negative semantic meaning, which it is not the case depicted in the map. A possible solution would be inverting the orientation of the colour scale, resulting in a blue-white-red scheme. Another possibility would be choosing a different diverging scheme, such as those proposed in the ColorBrewer page². In particular, it would be important to pick up a scale that is also understandable for people affected by some sort of colour-blindness problem.

The problem highlighted by this comment also applies to the case of the balloon plot used in the *Comparative overview* tab, as shown in Figure 8.

² <http://colorbrewer2.org/#type=diverging&scheme=BrBG&n=3>

- *From the interviewers' observations, the participants had problems in selecting the desired solution:*
This issue was in our opinion a combination of the misunderstanding of the “criteria prioritisation” selector, which could be replaced by “selection of criteria” or something similar; and the difficulty to highlight a solution in the PCP, which requires the user to point over the left column and not over an individual line.
- *The participants had also difficulties to deploy the tooltips of the clusters in the analysis window:*
It can happen that the tooltips require more space than the actual dimensions of the map window to be entirely deployed. This is particularly evident when text lines are long. A possible solution to this issue would be splitting the text across multiple lines, resulting in more compact tooltips. Another possibility is to provide some functionality to automatically compute the best place where deploying the tooltip, by taking into account the space available between the graphical object under analysis and its distance to the borders of the window. At the moment, the default behaviour is to show the tooltip always at the right side of the object.
- *Most users only required one chart (the PCP in the case of the solution selection and the “Comparison Overview” in the case of the solution analysis) to complete the exercise:*
This fact could mean that the general layout and workflow prevents users from exploring further options that could lead to better analyses.
- *Some users misunderstood the representation of the variable “fight_level_deviation” in the PCP:*
This variable represents the difference between the average flight levels employed in the represented solutions and the baseline scenario. Since 0 is the ideal situation, when the value of the variable is positive, a reduction means improvement while an increment denotes that the situation gets worse. The opposite is true when the value is negative. This is difficult to understand. It would be easier to interpret if only the magnitude of the variable were represented, since the direction of the deviation is not specifically meaningful.
- *A participant misunderstood the meaning of the variable unit rates:*
Due to the units of measure used to present the variable in the *Analysis window*, it is possible to intuitively understand the values as the total income obtained by the ANSPs instead of the rate received for the services provided at each unit of distance.

Presentation issues

- *When comparing the effects for a specific actor (e.g., airlines), both the Comparison Overview and the Detailed Comparison tabs should be set to show the percentage representation by default, as it would be more meaningful:*
We will take into account this suggestion when releasing an updated version of the dashboard.
- *Text and elements size are perceived in general as small:*
There is a compromise between the amount of information presented to the user and the available space in the browser. In general, text and elements have been reduced to avoid overlapping (especially when working with small screen monitors and dealing with a large number of indicators). This concern could be overcome by dynamically changing the text size based on the user selections and available blank space on the screen.

- *The labels identifying the routes in the map of the Analysis Window are confusing (perceived as measures):*

This issue may arise because these labels are numeric (as they come from the original data) and thus, people tend to misinterpret them as the numerical values of the variable represented. A possible solution would be to either remove them from the map or change them by including some alphabetical character to reinforce their label role (e.g., by naming the routes as r0, r1 and so on).

- *The quantitative information related to the comparisons presented in the Analysis Windows has not been correctly perceived as differences of the indicator effects between the selected solution and the baseline scenario:*

Some of the participants struggled to interpret correctly the meaning of the numbers placed in the balloon plot of the *Comparison overview* tab. We believe that one of the possible causes for this issue is the colour scheme used to represent the data, as already highlighted above. Another reason could be that it is not explicitly stated that the differences are computed by subtracting the baseline from the selected solution. In this sense, both the balloon plot title and the third radio button label in the map section seem to provide mistakeable information. In the next update, this issue will be tackled by rephrasing the ambiguous pieces of text.

Technical issues

- *Some interviewers experienced minor glitches using the tool (e.g., scatter plot not being rendered, headers of the views not being updated):*

Participants which have experienced such problems were using computers running MacOS or Ubuntu Linux. Despite the glitches were solved by refreshing the webpage in the navigator program, some more testing is required to understand the reasons behind these issues. Among the different reasons to build a web-based platform, there is the idea of providing a product/service that is platform-independent. However, it is not browser-independent, since it is well-known that browser compatibility is a great hurdle for web development. This is why we recommend to use a modern browser with a full HTML5 / CSS3 support (e.g., Firefox, Chrome, Opera) in order to limit the glitches described above. At the same time, it is possible that the problem was due to communication issues between the server and the client.

4 Conclusions and future work

In this document we have presented the methodology followed to get a comprehensive evaluation of the INTUIT platform developed in WP5 and the results obtained from the user assessment tests.

These preliminary results provide several indications about the main benefits and drawbacks found by potential users. Their analysis provides several insights about the improvements required to come with a more user-friendly version. At the same time, the hints collected throughout the evaluation phase bring new ideas about how to extend the dashboard functionalities in order to deal with a real operational environment.

The remarks collected show that the overall users' assessment is positive. Users believe that a wider adoption of such kind of visualisation tools at organisational level would be advantageous, especially when studying the trade-offs faced by specific categories of stakeholders (e.g., airlines). This is particularly important since the users involved are ATM experts at various degrees, and thus possible future users of similar visualisation tools.

On the other hand, some flaws have been detected during the evaluation experiment. These issues mainly concern the usability and/or the presentation of the content in some of the interactive items (that is, charts as well as input widgets) tested by the users. A small percentage of the issues highlighted by the testers may come from a lack of familiarity with the visualisation environment proposed (none of the volunteers participated neither in the design nor in the implementation steps; moreover, they played with the interactive environment only for about thirty minutes during the evaluation process). However, there are issues affecting user experience that should be tackled and worked out properly to provide a more effective and useful tool. Most concerns involve: the sliders used to set the criteria weights and filters, which were perceived as too similar; the colour scales encoding the comparison of the effects of the selected solution with respect to the baseline scenario; the lack of a user guide to help the user to understand how to use the different tools; and the labels and chart titles not providing appropriate information to interpret the results.

Finally, the volunteers highlighted that the trade-off analysis tools are among the best contributions provided by the dashboard, and the *Analysis window* was proved to be useful and adequate to present a detailed analysis of the impact of a specific solution on the different stakeholders involved.