D6.3 Visualisation Tools for Evidence-based Decision Support

Project acronym:	INSIGHT
Project title:	Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development
Grant Agreement number:	611307
Funding Scheme:	Collaborative project
Project start date / Duration:	01 Oct 2013 / 36 months
Call Topic:	FP7.ICT.2013-10
Project web-site:	http://www.insight-fp7.eu/
Deliverable:	D6.3 Visualisation Tools for Evidence-based Decision Support
Issue	1
Date	30/09/2016
Status	Approved
Dissemination Level:	Public

Authoring and Approval

Prepared by		
Name & Affiliation	**Position**	**Date**
Luca Piovano (UPM)	Researcher	23/09/2016

Reviewed by		
Name & Affiliation	**Position**	**Date**
Ricardo Herranz (Nommon)	Scientific/Technical Coordinator	30/09/2016
Iris Galloso (Nommon)	Management coordinator	30/09/2016

Approved for submission to the European Commission by		
Name & Affiliation	**Position**	**Date**
Ricardo Herranz (Nommon)	Scientific/Technical Coordinator	30/09/2016
Iris Galloso (Nommon)	Management Coordinator	30/09/2016

Record of Revisions

Edition	Date	Description/Changes
Issue 1 Draft 1	26/07/2016	First draft of the document
Issue 1 Draft 2	23/09/2016	Revised version of the document following the internal review of Issue 1 Draft 1; Modifications to the structure of Sections 2.2, 2.3 (sub-sections added) and 3 (content of subsection 3.1 moved to the main body of section 3);
Issue 1	30/09/2016	Minor editorial corrections and approval for delivery to the EC

Table of Contents

Executive summary

This document corresponds to the deliverable D6.3 of INSIGHT and describes the most relevant findings obtained through the work of the task WP6.3 – Visualisation Tools for Evidence-based Decision Support. The emphasis of this document is twofold: on the one hand, it details the main features, advantages and drawbacks of the polygonal scatterplot introduced to support the visual analysis of multi-dimensional data and, on the other, it presents practical evidence derived from the use of the visualisation platform to analyse the policy questions posed by the cordon toll measure proposed for Madrid. In particular, the main results presented in this deliverable are the following:

- An improved version of the polygonal scatterplot chart proposed by Chen et al. (2013), adapted to the requirements of the policy making process to be supported by the INSIGHT visualisation platform. In particular, our contribution encompasses: i) the definition of new point arrangement strategies aimed at providing more insights about the quality of trade-offs among different dimensions; ii) the adoption of new graphical elements to highlight specific features; and, iii) the addition of several, interactive features to allow a better exploration of the data represented;
- An integrated visualisation platform able to: i) play as a unified workspace for different simulators; ii) present scientific evidence in a friendly and flexible manner for a thorough policy comprehension; and, iii) set a bridge among disciplines and professionals in the fields of decision making and data analysis and simulation.

For the sake of completeness, we include also a discussion on the most important pros and cons of the polygonal scatterplot chart proposed, as well as some hints on how to improve it in the future. The discussion is illustrated with examples built using either simulation data or synthetic datasets.

Further improvements are required to bring the INSIGHT visualisation platform to the point where it can be used to support decision making in a real operational environment. Among the features that need to be implemented/strengthen are:

- Extending the library of charts to capture other data features / allow new data insights;
- Enhancing the analysis capabilities by providing suitable interfaces with statistical software (e.g. R);
- Expanding the number of simulators;
- Allowing specific connections between a simulator and the platform to build tailored services.

1. Introduction

In the context of the INSIGHT project, and specifically in WP6, one of the biggest challenges consisted in dealing with multiple simulation outputs characterised by a multi-dimensional (and heterogeneous) indicator space. The need to characterise the policy alternative outcomes through their indicator dimensions is essential in the context of the urban planning process, since it provides the measures of how well a policy formulation is expected to perform to solve urban issues. In this context, this essentially comprises two facets. First of all, it means comparing different policy formulations and detect possible trade-offs, drawbacks and strong points of each alternative along the dimensions of interest. Second, in the case a spatio-temporal information is associated to each alternative, it would be interesting to reason about both geographical and temporal patterns to spot the impacts on the territory of study. Unfortunately, many formulations comprise conflicting objectives to take into account, so that characterising the best (set of) solutions is indeed a difficult task that decision makers have to tackle.

Visualising and providing the right interfaces to explore the indicator space are among the objectives addressed in the definition of the visualisation platform detailed in *D6.1 – Visual Ecosystem Technical Specification and Design*. In this document, we describe our contribution to solve the multi-dimensional visualisation problem by describing the polygonal scatterplot charts. Starting from the original proposal by Chen et al. (2013), we elaborated advances in both the approach and the graphical elements for the exploration and understanding of the indicator space. In particular, we mainly focus on the point arrangement strategies and the interactive capabilities to get more details on the policy formulations. The objective is to highlight the main research contributions as well as to provide a comprehensive guide for potential users being interested in adopting our methodology (see Annex I for related definitions). This chart has been inserted in our visualisation platform as a component stimulating the interpretation and communication of policy alternative results by enhancing the analytical reasoning abilities of users. In this sense, *D6.2 – Visualisation Tools for Simulation and Scenario Analysis* has already introduced some of the (operational) aspects covered here, but with the purpose of providing a general picture of the capabilities of the platform through the description of its main components, their interactions and general functionalities. The following document completes the series of the deliverables for WP6 by focusing on the new tools for evidence-based decision support and impact assessment contributed by INSIGHT, as well as on their application to the case studies defined by the project. Indeed, the last part of the work focuses on the description of a possible example of use of the visualisation platform, related to the cordon toll policy for the city of Madrid, with the aim to highlight its functionalities in an operational context.

Therefore, this deliverable is organised as follows:

- **Chapter 1** provides the overall frame for WP6.3;
- **Chapter 2** introduces a new approach to represent multi-dimensional data based on the polygonal scatterplot proposed by Chen et al. (2013). In this section, we outline the main features of this chart and the differences between the original proposal and our implementation;
- **Chapter 3** presents examples of application to relevant case studies developed by INSIGHT; and,
- **Chapter 4** highlights the main contributions put in place by the INSIGHT visualisation platform, discusses related challenges and potential improvements and introduces future research avenues.

2. Polygonal scatterplot

One of the biggest challenges in the decision making process is to deal with policy proposals addressing multiple, conflicting objectives (see Annex I for related definitions). In particular, if the policy options space is considerably large, the hardest task is to find the best candidate solution(s) solving the policy problem. Exploring such an objective space – which could be formally defined as a mapping of each possible candidate solution to the values of the multiple-objective function under consideration - is not a trivial task, especially when more than three objectives are considered at the same time. One of the key elements to take into account in such a process is the trade-off evaluation of the different alternatives, provided that it is really unlikely to find a solution being able to optimise each objective at the same time (i.e. without having a negative impact on the other dimensions). In such a conflicting situation, the resulting set of solutions is called Pareto optimal frontier and could possibly count infinite solutions. In this context, the meaning of the word 'optimal' has strong subjective components and largely depends on the point(s) of view the analyst intends to focus on.

Effective visualisation approaches have to deal with such issues in order to both encourage the interactive exploration and selection of the best alternatives and provide strong and defensible evidence to strengthen the whole decision making process. Possible goals to achieve comprise: i) finding a set of Pareto optimal solutions with respect to a specific set of objectives; ii) allowing the user to use suitable filters to narrow the search space and adapt it to his/her preferences; and iii) quantifying trade-offs. In all these cases, the key goal of the visualisation deals with finding a proper way to visually represent a multi-dimensional space. In the field of multi-criteria decision making (MCDM in the following), several approaches have been proposed to the date. A couple of interesting reviews could help to introduce the discussion on the most common techniques used in representing solutions of a multi-objective optimisation problem (see (Korhonen & Wallenius, 2008) for visualisation in MCDM frameworks and (Lotov & Miettinen, 2008) for Pareto frontiers). Some of them comprise, but are not limited to, radar charts, Chernoff's faces, parallel coordinate plots and their smoothened version called Andrews' curves, and interactive decision maps.

However, each of these approaches present some inadequacy when put in practice, as described in **¡Error! No se encuentra el origen de la referencia.** below. The need to refine the current visualisation techniques is getting more and more urgent in light of the increasing interest in embedding visual analytics approaches in the current decision making process steps. In this sense, one of the most promising works dealing with this challenge is the one by Chen and colleagues (Chen et al., 2013), which describes a visual methodology based on a modification of the Self-Organizing Map (SOM) approach (Kohonen, 1990). Starting from that work, we embedded into our visualisation platform an adjusted version of the original SOMMOS (Self-Organizing Maps for Multiple ObjectiveS) chart to intercept the multi-dimensional features of the simulation data. More precisely, in the analysis of alternative policy options, the polygonal scatterplot is used with a twofold aim: to depict possible trade-offs of the alternative policy options under a restricted number of dimensions / objectives and, to analyse how a specific policy performs across the different urban areas of study (for instance, the postal code regions in Rotterdam or the census sections in Barcelona) given a subset of dimensions / objectives; and then, it could be used to characterise how a single policy could perform across the different urban areas of study (for instance, the postal code regions of Rotterdam or the census sections in Barcelona) given a subset of dimensions / objectives. This latter point incidentally shows a possible new use of this chart with respect to the original formulation which was limited to the former case instead.

2.1 Chart anatomy

The polygonal scatterplot was conceived as a mean to combine the strengths of radar charts and traditional scatterplots in order to extend their capabilities, overcome their issues, and last but not least, show data relationships and trade-offs in a 'multi-dimensional' space. An example of such a chart is provided in Figure 1.

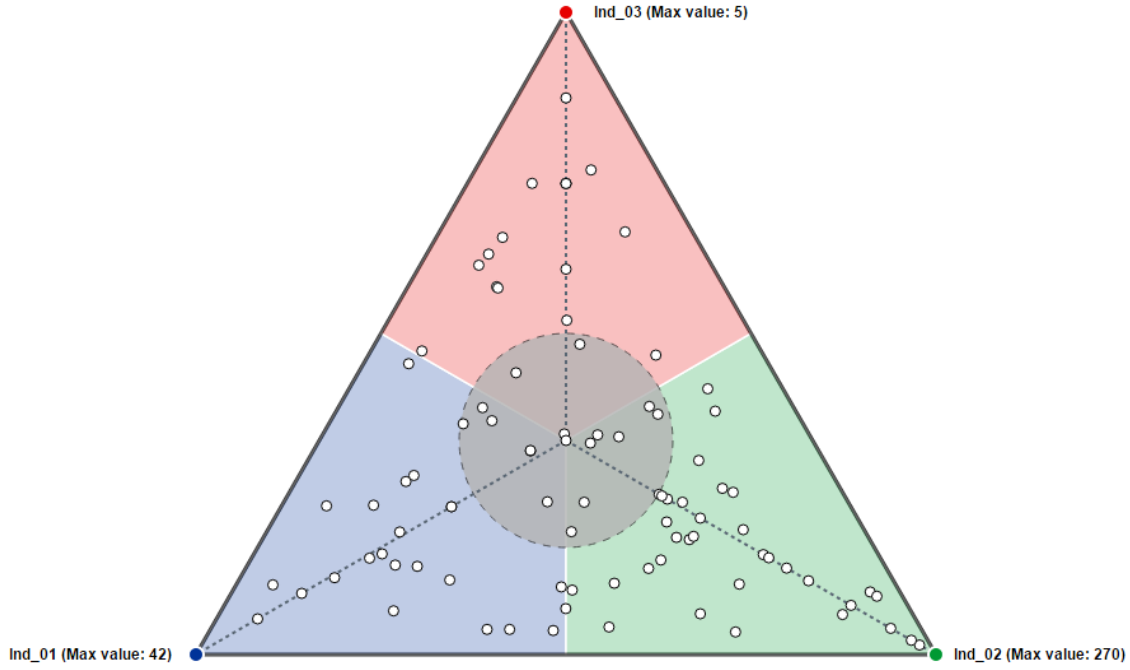

Figure 1 – Triangle-shaped scatterplot and its main components (point arrangement strategy: weighted barycentre)

It basically consists of an n-gon in which a set of points is drawn, being n ≥ 3 to properly draw a plane figure. Each vertex of the polygon represents an objective / dimension of the multi-dimensional space defining a given policy. A suitable label (as well as some additional information such as the maximum value for a given variable) is placed close to the vertex itself to provide more context to the user. Moreover, each vertex is easily recognizable because of its colour. On the other hand, each point is a specific data point measured in that n-dimensional space. In the context of our visualization platform, a point could be interpreted in two different ways, namely either as a policy alternative or as an urban area (following the territorial division defined by each simulator considered in the project). In the first case, the user aims at exploring and understanding the trade-offs among different policy options according to the n objectives previously selected. In the latter case, the user can explore the characterization of geographical areas according to a single policy alternative (as usual, limited to its n dimension previously selected). Whatever the aim of the analysis may be, the chart allows any user to support operational tasks such as comparison, search for patterns and relations, and locate groups and clusters of interest.

Other three features complete the definition of the polygonal scatterplot – namely, the radial axes, the Voronoi cells, and the sink – to foster the user's analytical reasoning. Each axis originates from the very centre of the polygon and connects this point to each polygon vertex in turn. Being the polygon built as a regular shape, all the axes have the same length. As in the case of a traditional Cartesian space, the axis defines both a dimension and its scale values, but in contrast with it, the axis of the polygonal scatterplot are always defined in the positive range$[0, +\infty]$. Since policy dimensions are very likely to be heterogeneous, the measures – in terms of both units and ranges – used to express them will too. In general, it is possible to say that:

$$\forall d_i, val(d_i) \in [min_{d_i}, max_{d_i}], \qquad -\infty \leq min_{d_i} \leq max_{d_i} \leq +\infty,$$

where di is the i-th dimension and val(di) is its corresponding measure value. In such a situation, comparisons across dimensions could be flawed and a step of normalization is often required. In our case, we map each dimension to the range $[0, 1]$. This way, we shift the attention of the user on the proportional value (with respect to the maximum value obtained for the i-th dimension) of each component rather than on its raw values. It is worth to clarify at this point that we work under the assumption that the analysed objectives / dimensions can be quantitatively expressed. That is, we are not considering here qualitative variables.

Another component of the chart anatomy is the set of Voronoi cells tessellating the inner space of the polygon. In our approach, they are generated considering the polygon vertexes as the seeds: as a consequence, the cells have equal sizes and areas. To facilitate their recognition, they are coloured as the corresponding vertex (adding a little bit of transparency to enhance the visual contrast against the white points inside). From an interpretative point of view, these Voronoi polygons, in combination with the point locations, define the area of influence of a given dimension, or, said in a different manner, they help the user to detect which set of points have a greater contribution from that dimension. An example might help to understand the point. With respect to Figure 1, and limiting our analysis to the red zone of the indicator 'Ind_03', it is possible to make assertions like the following:

- just a small fraction of the points lays inside the red area (that is, not many points have that indicator as the dominant dimension);
- some of them lay really close to the borders with either the blue and green zones, indicating that there is a sort of balance between the corresponding components;
- some of them lay on the red axis: the closer to the red vertex, the closer to zero the contributions of the blue and green dimensions; on the other hand, the closer to the sink area, the more similar the contributions of each dimension (i.e. the dimensions tend to an equilibrium);
- some of them are much closer to the triangle centre, meaning that they have globally well-balanced contributions from each dimension);
- the remaining points are spread through the red polygon: if they lay on the leftmost (respectively rightmost) side of the axis, they reveal a more or less intense interaction with the blue (respectively green) dimension, depending on its closeness to the corresponding vertex.

As observed by this example, the closer a point to a vertex, the stronger the contribution carried by that dimension. We call this property orientation and we will discuss it with more details later on in this section. For the moment, it is important to note that there could be some exception to this rule, that is a point with a strong value for a given dimension (say the red one, to continue with the previous example) would not be placed close to the corresponding vertex (the red one) but somewhere else. Since the idea of the polygonal scatterplot is to map the relative contributions of each dimension, this fact could happen whenever another dimension (say the green one) exceeds its overall impact contribution on a specific point, so that it would be correctly placed more or less close to the stronger vertex (the green one). Detecting these cases is impossible by just looking at the situation as depicted in Figure 1. To overcome this issue, a combination of interactive exploration of the polygonal space and changing the placement schemes are valuable supports to correctly interpret the situation depicted in the chart.

Last, the sink (or equilibrium area) is another useful feature used to carry to the user's perception more information on what (s)he is currently seeing. In Figure 1, this element is depicted as the grey circle at the very

centre of the polygon and intersecting the three axes. Its main function deals with highlighting the zone where points are in equilibrium with respect to the represented dimensions. A policy alternative / urban area is said to be in equilibrium if the contributions of all the dimensions under analysis are really similar to a greater or lesser extent. It is important to note that the equilibrium condition is reached regardless of the actual values across the dimensions: what really matters is that values have to differ little among each other. The closer such a difference to zero, the closer the point to the polygon centre. As a consequence, the sink area could be packed with both good alternatives (i.e. scoring good / optimal measure values in the dimensions analysed) and bad options (with very little contribution of all the dimensions involved). Since it is not possible to clearly distinguish the two cases (a clear flaw of this representation), the best option to overcome it consists in filtering out the worst alternatives. In this sense, a good strategy would be to prevent to show those points whose entirety of dimension values do not exceed a given threshold, such as, for instance, the value intercepted at the axes (i.e. the sink radius). In Figure 1 this value is set to 0.25 and if the filter would be activated, the points scoring less than 0.25 across all its dimensions would not appear in the chart. The smaller the circle, the less the likelihood for points to fall inside it, meaning that the equilibrium condition is more restrictive. Conversely, a larger sink could take over too many points resulting in possible mismatches in labelling policies or areas as in equilibrium. With respect to the point arrangement strategy (see Section 2.2), the information told by the sink is more informative when a linear representation is chosen (e.g. weighted barycentre or top-2 dimensions); in all the other cases, the equilibrium condition is really hard to represent and as such, the sink element could be turned off. However, since the default arrangement strategy is the weighted barycentre, the sink is always turned on (as it could be seen in all the Figures in this chapter).

2.2 Points arrangement strategies

One of the key issue in drawing a polygonal scatterplot chart is finding a suitable strategy to representatively place points inside the shape such that true patterns and relationships hidden in the data could emerge. This task is really a troublesome situation because, regardless of the followed approach, it concerns performing a mapping from $\mathbb{R}$m $\rightarrow \mathbb{R}2$, that is to say passing form a multi-dimensional to a bi-dimensional domain. From a mathematical point of view, this operation inherently carries a loss of information, as we are projecting on a flat surface more than 2 dimensions. This way, the spatial relationships across multiple dimensions may be lost. For that end, interactive exploration of different placement strategies supported by subsequent selections of overlapping sets of objectives (in a way that at least one dimension is always present in different selections) could be a first approach mitigating such an issue. At the same time, presenting different polygonal charts according to the small multiple approach could be another effective way to help the user to foster his/her analytical reasoning about different partitions of the data under analysis.

For a sake of readability, let us provide some object definitions used throughout this and the following sections.

A data point d is an m-dimensional vector defining a policy / region across m pre-defined dimensions, being $m \leq n$, where n is the total number of the original dimensions considered to fully characterise the policy / region under analysis and m corresponds to the number of sides of the polygon. This way:

$$d = [v_1, v_2, \cdots, v_m],$$

where each v_i is a vector component describing the (normalised) value of the *i*-th dimension.

On the other side, a chart point p is a bi-dimensional vector describing the *(x,y)*-coordinates inside the polygonal space where to draw the point itself.

A point arrangement strategy / algorithm is a function f, such that:

$$f(d_i) = p_i, \qquad f: \mathbb{R}^m \rightarrow \mathbb{R}^2$$

All the points arrangement algorithms described in the remainder of this section have in common a two-fold, high-level strategy:

- Each dimension exerts an influence (force) on the point based on its corresponding value and the final position of such a point is given by a combination of such influences;
- Points tend to be placed in the polygonal space such that the conflicting properties of adjacency and orientation – as defined in (Chen et al., 2013) – would hold. These properties translate into two main general principles. The principle of adjacency indicates that the more similar the multi-dimensional data vectors describing a policy object are, the closer the corresponding points may be placed. Conversely, orientation establishes that points should be placed as close as possible to the vertexes (i.e. dimensions) that are (relatively) dominant.

We have currently tested five different strategies, namely weighted barycentre, top-2 dimensions, dimensions sieve, similarity enhancement and data fit. In the remaining of this section we will briefly discuss each of them.

2.2.1 Weighted barycenter

The most straightforward strategy to place a data point inside a polygon is to consider the mean of its vector dimensions, that is:

$$p_{i,x} = \left(v_{1,x} + v_{2,x} + \cdots + v_{m,x}\right)/m,$$
$$p_{i,y} = \left(v_{1,y} + v_{2,y} + \cdots + v_{m,y}\right)/m,$$

where the sub-indexes *i,x* and *i,y* represent the *i*-th projections of a component with respect to the x and y axes respectively. However, this strategy tends to group points close to the sink, resulting in a poor informative representation and this trends accentuates more and more if the number of dimensions *m* increases. To overcome the issue, we modify this approach to weight the normalised values of the data vectors in a way that more dominant dimensions result with a higher force applied on the point. For this end, we consider the data vector as split in 4 different, equally-spaced ranges (i.e. each of these has a length of 0.25) and assign to them a corresponding weight value. Thus, the point coordinates are computed as:

$$p_{i,x} = \frac{w_1}{|r_1|} * \sum_{v_{i,x} \in r_1} v_{i,x} + \frac{w_2}{|r_2|} * \sum_{v_{i,x} \in r_2} v_{i,x} + \frac{w_3}{|r_3|} * \sum_{v_{i,x} \in r_3} v_{i,x} + \frac{w_4}{|r_4|} * \sum_{v_{i,x} \in r_4} v_{i,x},$$
$$p_{i,y} = \frac{w_1}{|r_1|} * \sum_{v_{i,y} \in r_1} v_{i,y} + \frac{w_2}{|r_2|} * \sum_{v_{i,y} \in r_2} v_{i,y} + \frac{w_3}{|r_3|} * \sum_{v_{i,y} \in r_3} v_{i,y} + \frac{w_4}{|r_4|} * \sum_{v_{i,y} \in r_4} v_{i,y},$$

where r_i are the ranges the domain is divided into (i.e. $r_1 = [1 \ldots 0.75\ ]$, $r_2 = (0.75, \ldots 0.5]$, $r_3 = (0.5, \ldots, 0.25]$, and $r_4 = (0.25, \ldots, 0]$), $|r_i|$ is the cardinality of each of such ranges, and wi is the corresponding weight (e.g. $w_1 = 0.5$ $w_2 = 0.3$, $w_3 = 0.2$, and $w_4 = 0$ such that $\sum_i w_i = 1$.

Figure 2, Figure 3 and Figure 4 show some examples of such a strategy applied to triangular, squared and pentagonal shapes respectively. The set of points depicted in there is the same throughout the series of images illustrating the different arrangement strategies (i.e. until Figure 16).

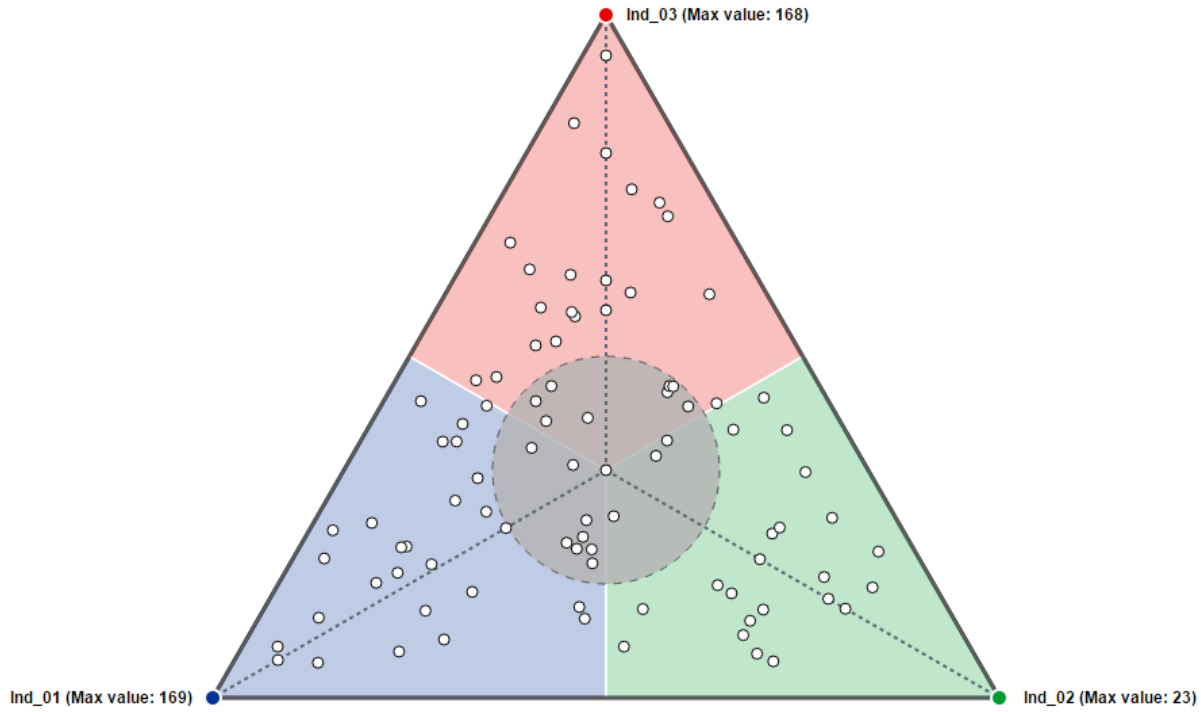

Figure 2 - A triangular scatterplot: points are arranged according to the weighted barycentre strategy

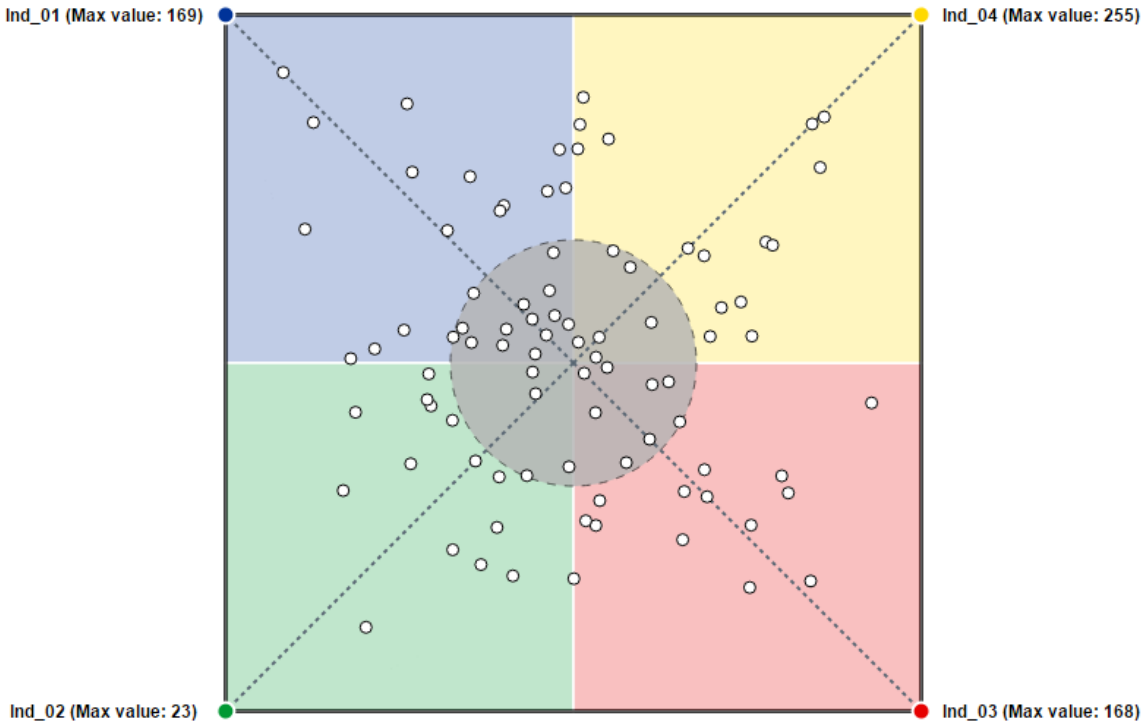

Figure 3 - A squared scatterplot: points are arranged according to the weighted barycentre strategy. Two facts outstand: many points lay inside the sink area and the overall distributions of points seems to be attracted by the centre of the chart. This is not surprising with this placement strategy since the forces exerting on the majority of points have contrasting effects and therefore tends to reach an equilibrium (i.e. levelling the effects of peaks). Nonetheless, some 'outliers' seems to appear revealing the underlying distributions among the dimension values. It is clear from this example that for a number of dimensions greater than 4, this placement strategy could limit someway the interpretative power of the chart. For this reason, other arrangements are required to unveil data properties in a greater detail.

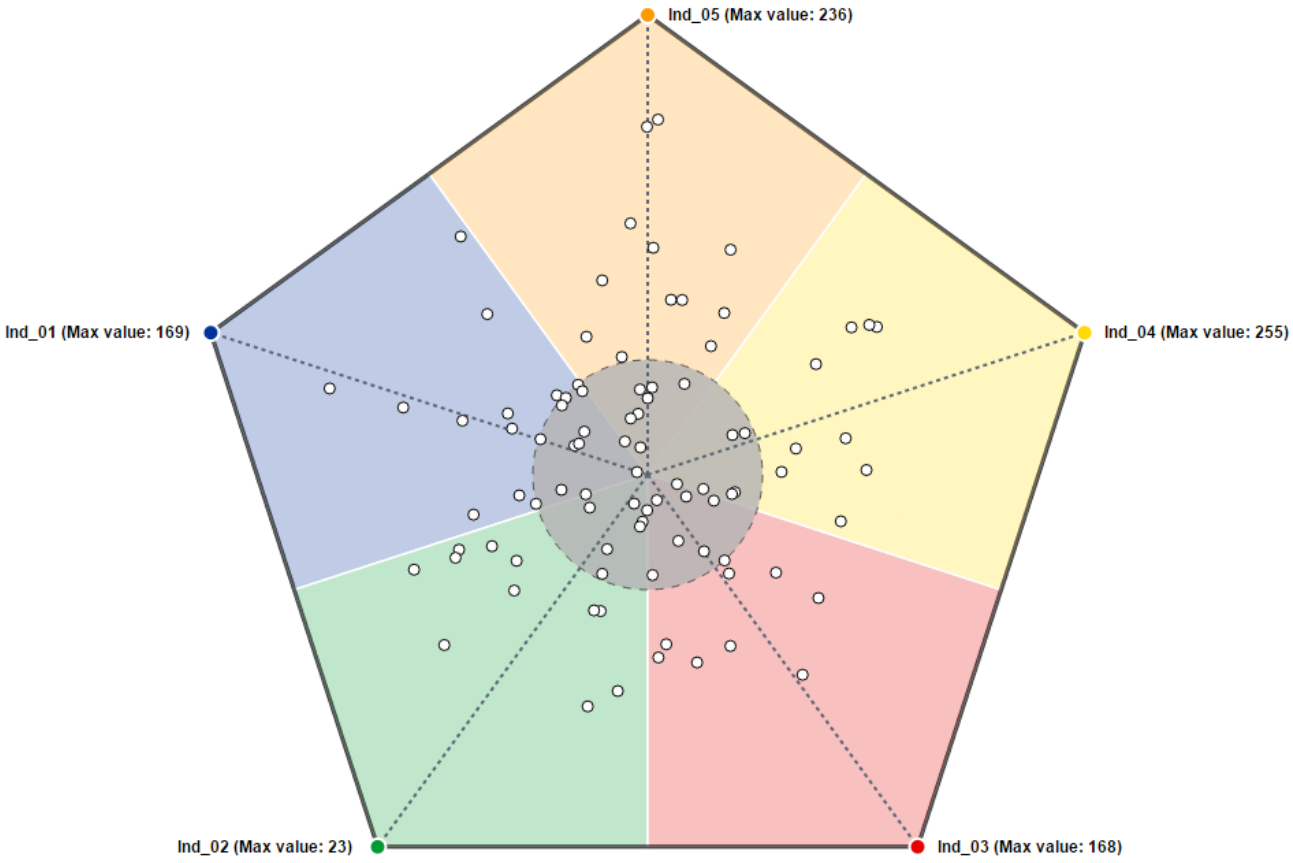

Figure 4 - A pentagonal scatterplot: points are arranged according to the weighted barycentre strategy. Effects of opposing forces are getting more and more evident since points tend to be represented inside the sink.

2.2.2 Top-2 dimensions

This is another modified version of the barycentre approach, such that just the two first dimensions with the higher values actively exert an attractive force on the point. In mathematical terms:

$$p_{i,x} = \left(v_{i,x}^1 + v_{i,x}^2\right) / 2,$$

$$p_{i,y} = \left(v_{i,y}^1 + v_{i,y}^2\right) / 2,$$

where the super-indexes 1 and 2 are not powers rather than the position of the component in the ordered vector.

This strategy allows to highlight which dimensions play the role of top-contributors for either a given policy alternative or an urban region. In both cases, they can suggest to the analyst in which areas the main effects are foreseen and, conversely, where it could be necessary to intervene to improve the current situation. From a visual point of view, points are usually found close to the edges joining the top-two dimensions and at the borders of the corresponding Voronoi cells. When showing a triangular scatterplot, points are usually displayed far from the sink too. However, in the case of dealing with more than 3 dimensions, some of them could lay close to its borders or inside it, because the vertexes of the top-contributor dimensions are located at the opposite of the figure.

Examples of triangular, squared, and pentagonal scatterplots are shown in Figure 5, Figure 6 and Figure 7.

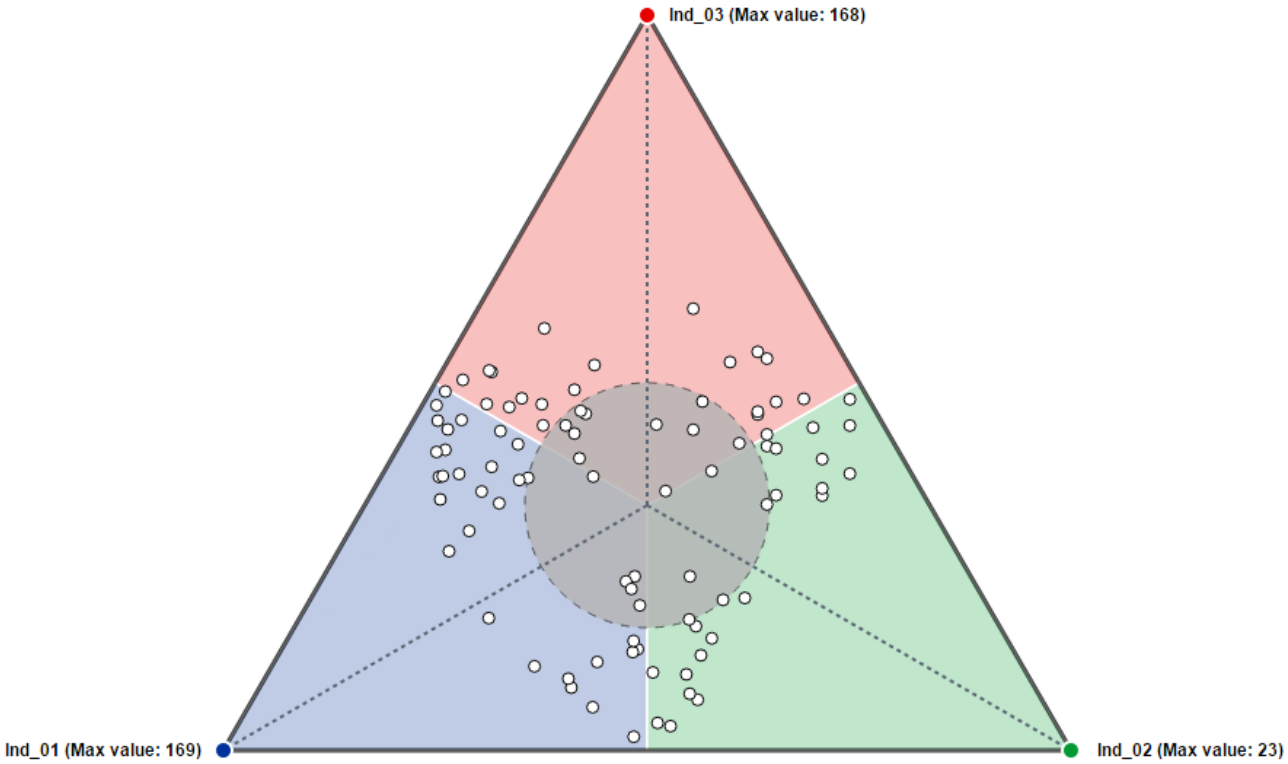

Figure 5 - A triangular scatterplot: points are arranged according to the top-2 dimensions strategy

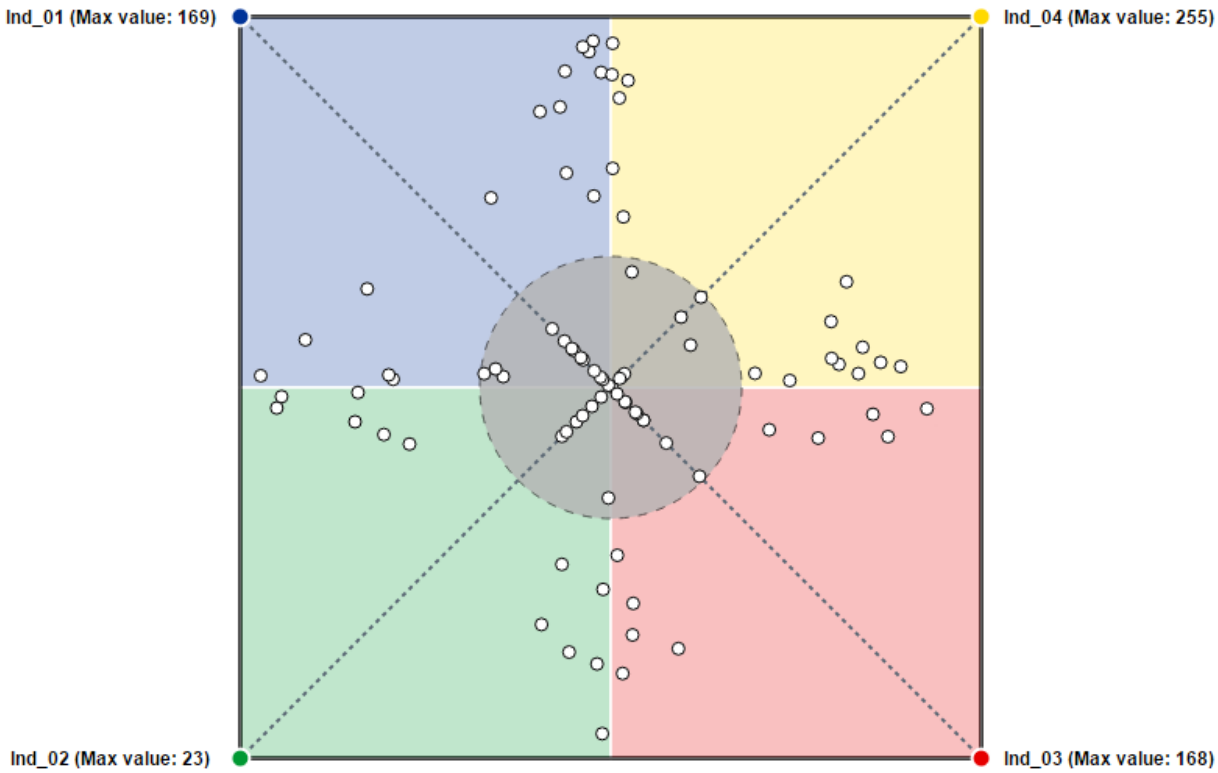

Figure 6 - A squared scatterplot: points are arranged according to the top-2 dimensions strategy. Points laying along the main axes of the square show how points are placed when forces are exerted by diametrically opposed vertexes. It is easy to see that points tend to stay close to the cells boundary lines. The cell where a point lays tells the user which is the most dominant dimension. By interacting with points, it is possible to discover more numerical properties of each point.

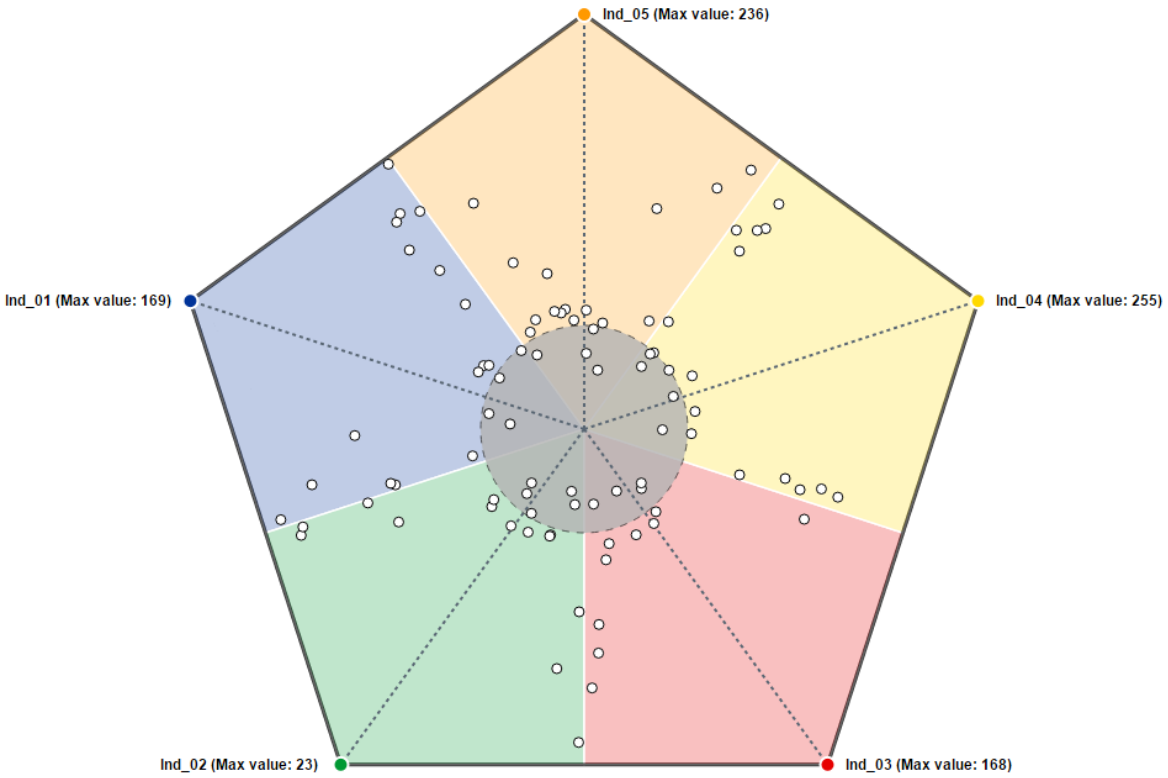

Figure 7 - A pentagonal scatterplot: points are arranged according to the top-2 dimensions strategy. For a number of dimensions n≥ 4, this approach could provide some misleading interpretation. For instance, consider the points in the green area and close to the sink: their position could suggest that they have small values for almost every component, but the green dimension at some extent. Instead, they represent data whose blue and red dimensions are the most prominent ones.

2.2.3 Dimensions sieve

So far we have just considered linear functions to compute the coordinates of the points to insert in the polygonal space. What about if the contributions would follow a different scheme? Of course the rationale is always the same, namely to promote top-level contributors and mitigate at the same time lower influences, such as it is possible to display in the chart which kind of trade-offs are necessary among the different dimensions. One of such approaches considers decaying laws to weight the contributions of each dimension. In particular, in the case of a time-decayed approach, we can iteratively apply a formulation as follows, provided that, to correctly apply this algorithm, we need to sort the dimension values in a descending order before:

$$p_j = (\sum_{i=1}^{m} \alpha * res_i * v_{j,x}^i, \sum_{i=1}^{m} \alpha * res_i * v_{j,y}^i), \qquad res_1 = 1, res_{i+1} = (1-\alpha) * res_i,$$

where $0 \leq \alpha \leq 1$ is the coefficient of decay and res_i represents the weighted fraction of terms to be considered in future iterations. Practically speaking, the rationale of this technique – adapted from a formulation widely used in several computer science applications, such as estimating future CPU burst time – is to assign different weights to the dimensions already considered with respect to the ones still left. Since the vectors are sorted in a decreasing order, the former case deals with those dimensions having the highest contributions and whose effects we wanted to visually amplify. For this end, it is important to assign a coefficient α such that $0 < \alpha < 0.5$. For instance, setting α to 0.3, the series of coefficients $c_i = \alpha * res_i$ in the sum terms of the first equation above will be:

$$c_1 = 0.3;\ c_2 = 0.21; c_3 = 0.147; c_4 = 0.1029; \ldots$$

showing how the importance of smaller dimension values is getting lower and lower.

As shown in Figure 8, Figure 9 and Figure 10, this arrangement strategy tends to form elongated clusters near the axes. This way it is possible to spot both the dominant dimension (one of the extremes of these clusters points towards a vertex of the polygon) and the second most relevant contribution (by looking at which half of the Voronoi cells the points lay in).

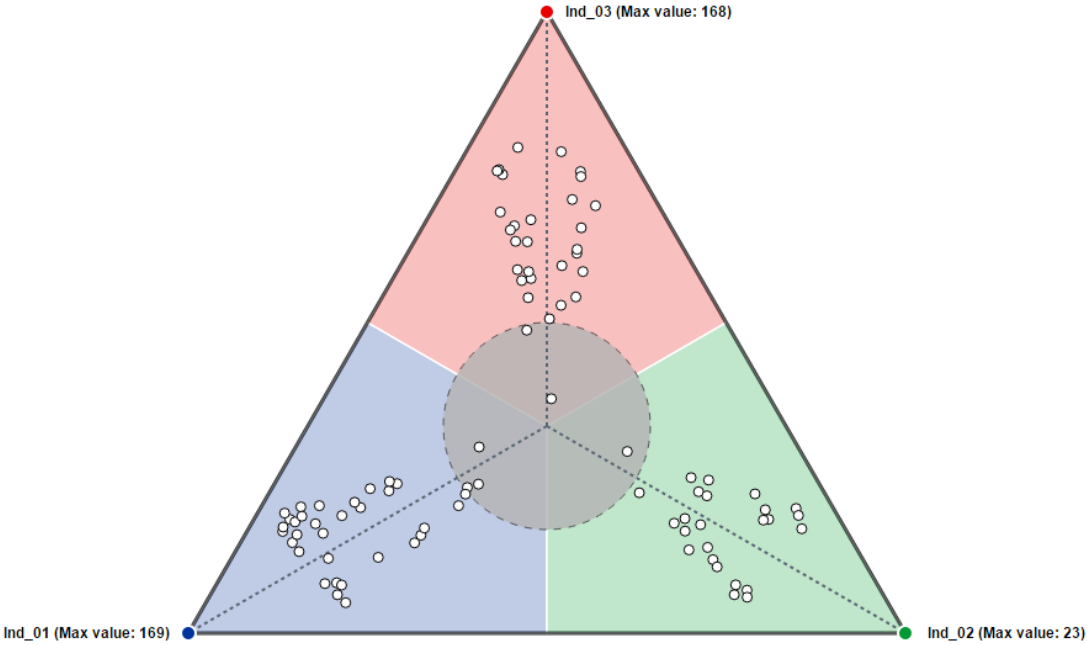

Figure 8 - A triangular scatterplot: points are arranged according to the dimensions sieve strategy

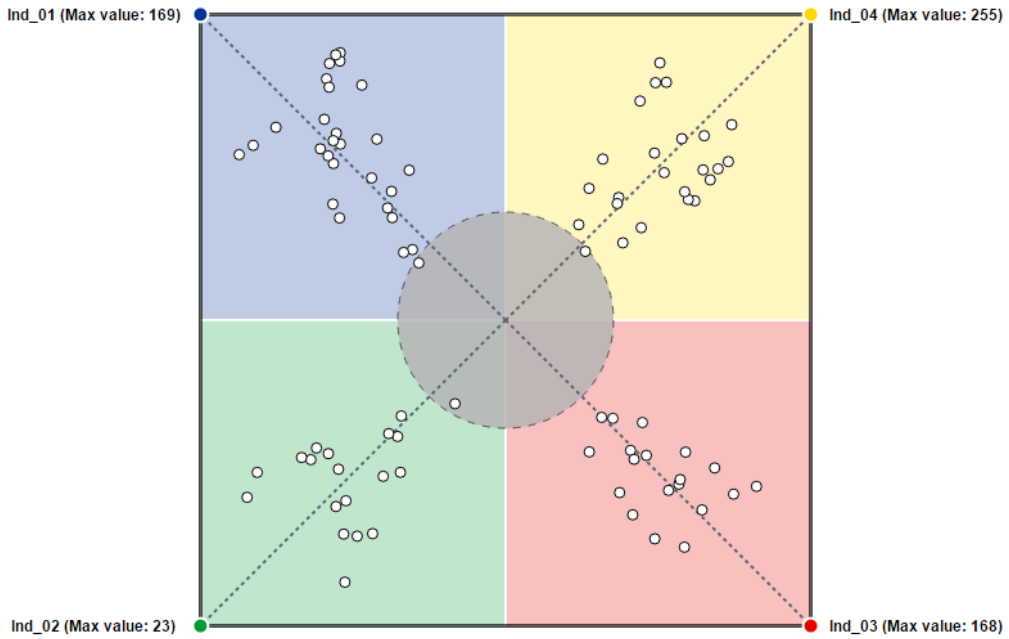

Figure 9 - A squared scatterplot: points are arranged according to the dimensions sieve strategy. It is possible to see, for instance, that the blue and yellow dimensions are the most dominant dimensions in the majority of the case. At the same time, points towards the centre of the chart are likely to have similar contributions either across two (but placed at the opposite corners) or three different dimensions.

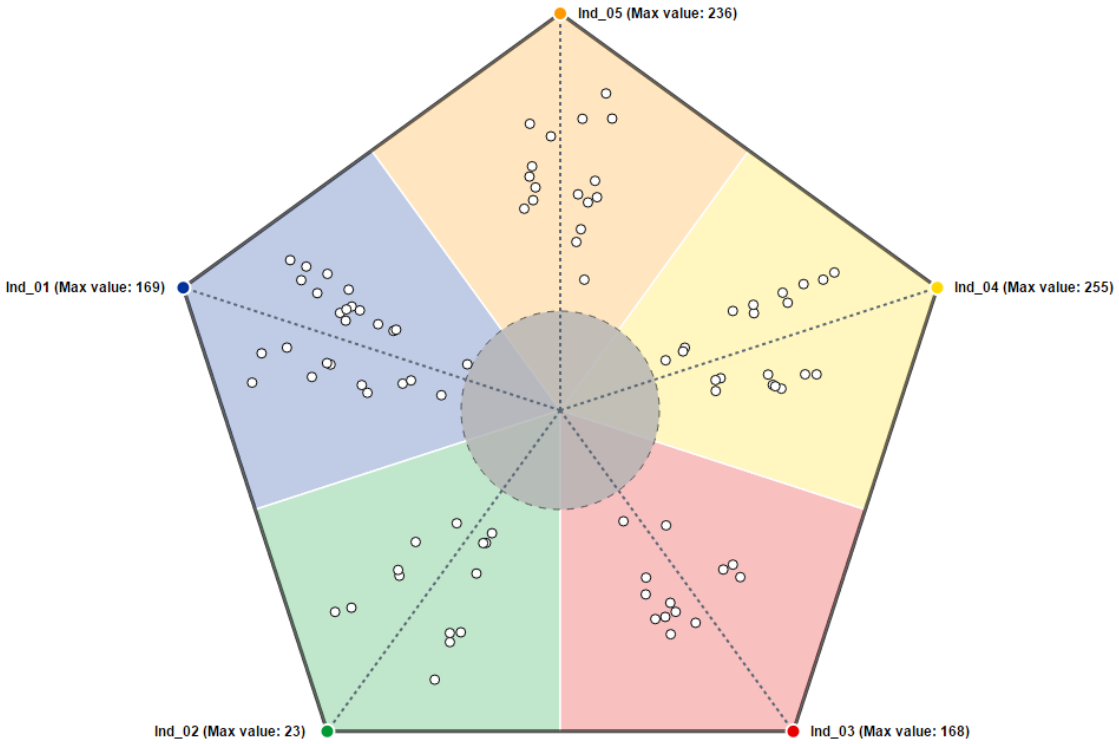

Figure 10 - A pentagonal scatterplot: points are arranged according to the dimensions sieve strategy. This representation is more robust against the problems highlighted in the previous figure.

2.2.4 Similarity enhancement

Figure 11, Figure 12 and Figure 13 show examples of the similarity enhancement strategy, which is a modified version of the previous approach.

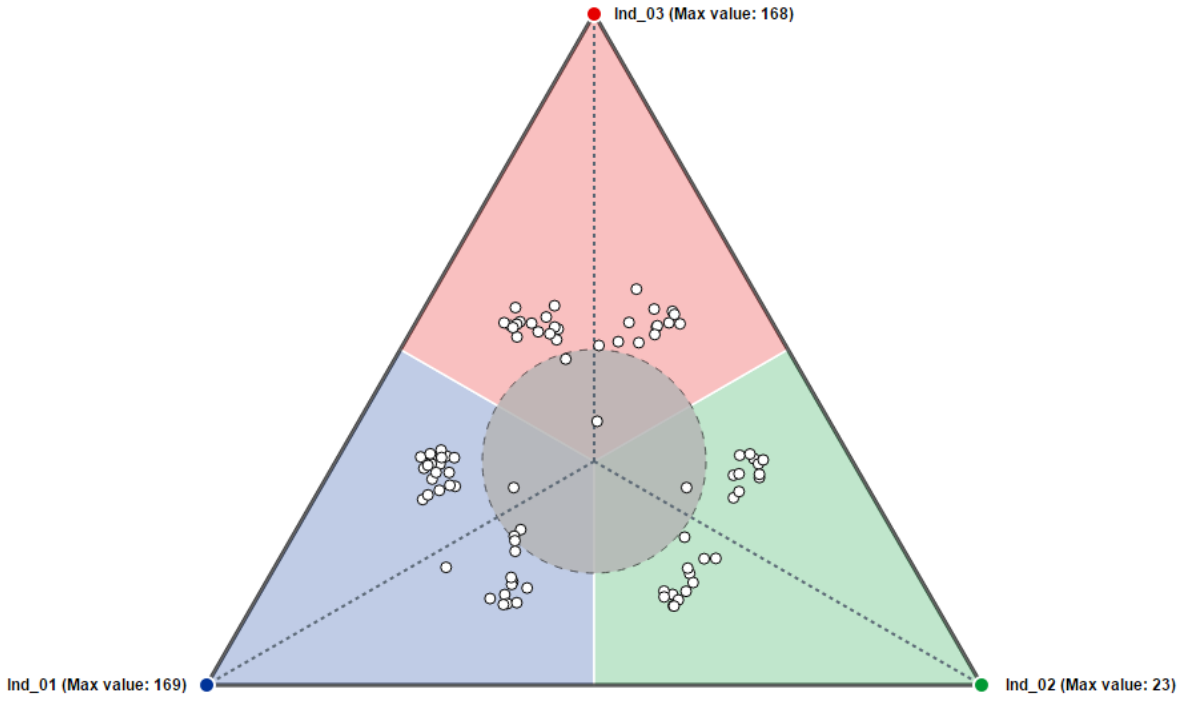

Figure 11 - A triangular scatterplot: points are arranged according to the similarity enhancement strategy

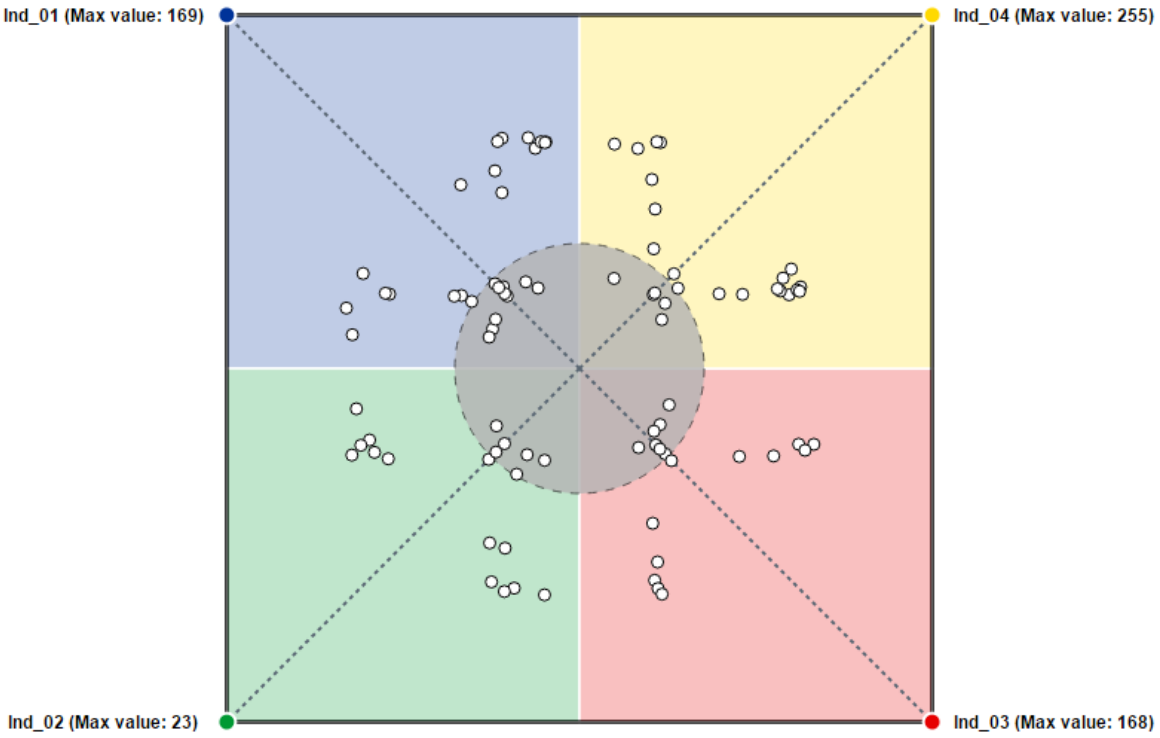

Figure 12 - A squared scatterplot: points are arranged according to the similarity enhancement strategy. It is evident the tendency to clustering revealing associations among different data points.

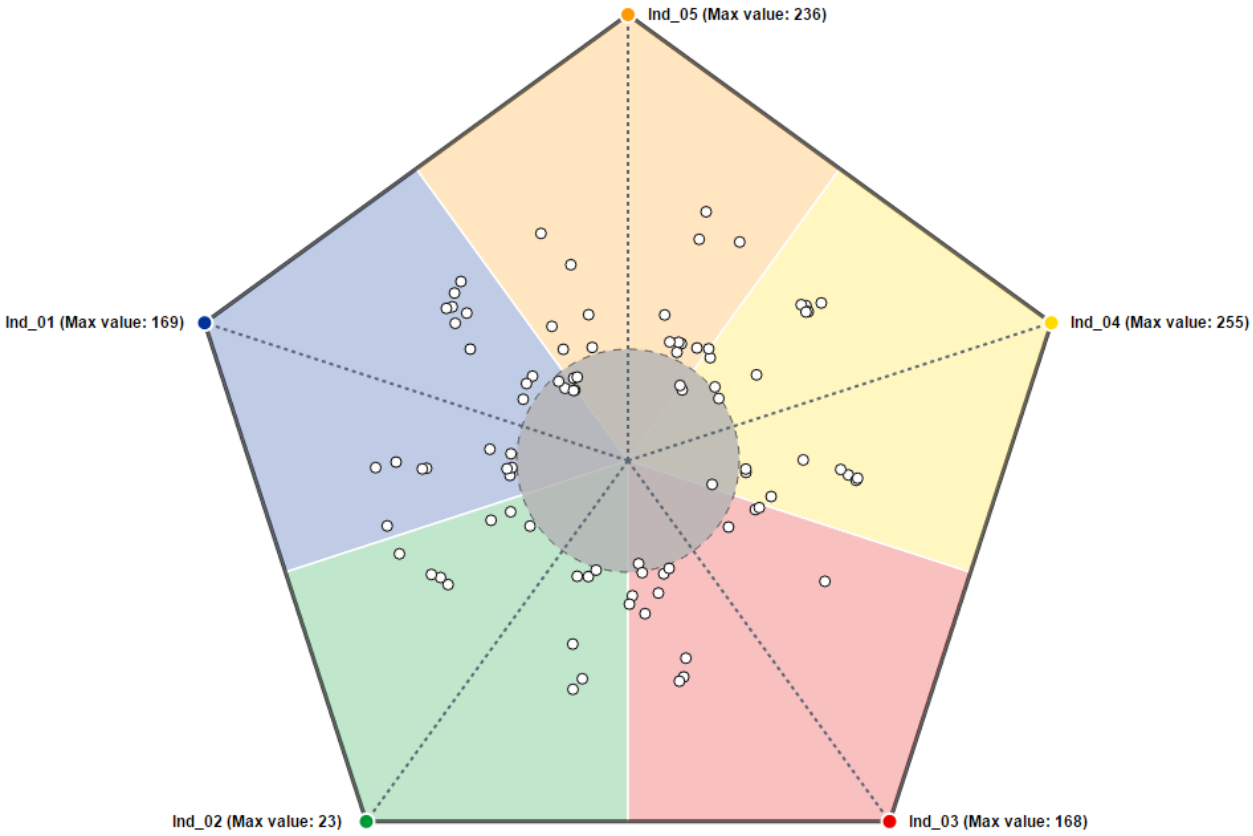

Figure 13 - A pentagonal scatterplot: points are arranged according to the similarity enhancement strategy. This approach tends to have the same flaws of the top-2 dimension strategy for a number of dimensions greater than 4.

The equations of reference in this case are as follows:

$$p_j = (\sum_{i=1}^{m} e^{-\lambda v_{j,x}^i} * res_i * v_{j,x}^i, \sum_{i=1}^{m} e^{-\lambda v_{j,y}^i} * res_i * v_{j,y}^i), \qquad res_1 = 1, res_{i+1} = res_i - e^{-\lambda v_j^i},$$

where λ is a constant tuning how fast the exponential decay tends to zero, the data vector has its components in descending order and the *i*-th component explicitly appears in the computation of the residual at the next step. Even with this approach, it is possible to spot clusters of data points: in this case, they are usually more compact and they can either provide information about the average effects of a policy in a group of regions or find an average alternative among different options performing more or less the same.

2.2.5 Data-fit

All the approaches described so far present some benefits and drawbacks when applied in practice because they have been designed to put in evidence just some data features rather than others. Changing the representation method and comparing the results is a good option to get a full picture of the data landscape under analysis. On the other hand, the setback is that a user can perform it only in consecutive steps with subsequent representations continually overwritten: as a consequence, the comparison power is reduced and the effectiveness of these methods lowered. To preserve space in the dashboard (i.e. to avoid the use of the small multiple approach) and enhance the comparison of different data features, we explored the possibility to have a data representation that could reveal the most of the features hidden in data values by analysing the value distributions themselves. The data-fit approach described here starts from this point: it tries to assign the best representation as possible to a data vector (among the ones previously described) by examining how its values are distributed across the different dimensions (see Figure 14, Figure 15 and Figure 16).

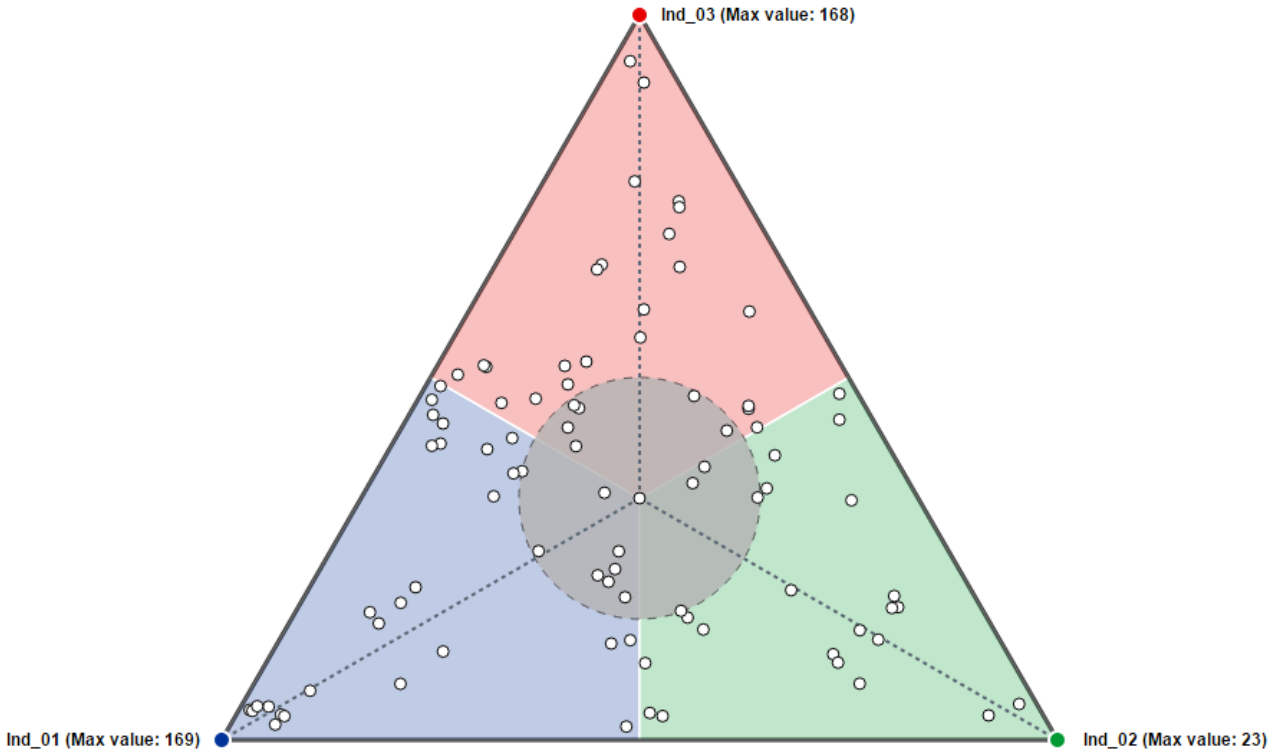

Figure 14 - A triangular scatterplot: points are arranged according to the data-fit strategy. At a first sight, it is possible to spot which data points have a higher contribution from a single dimension (i.e. the points closer to the corresponding vertex and almost laying on the corresponding axis) or two dimensions (i.e. the points laying close to the division line between two zones).

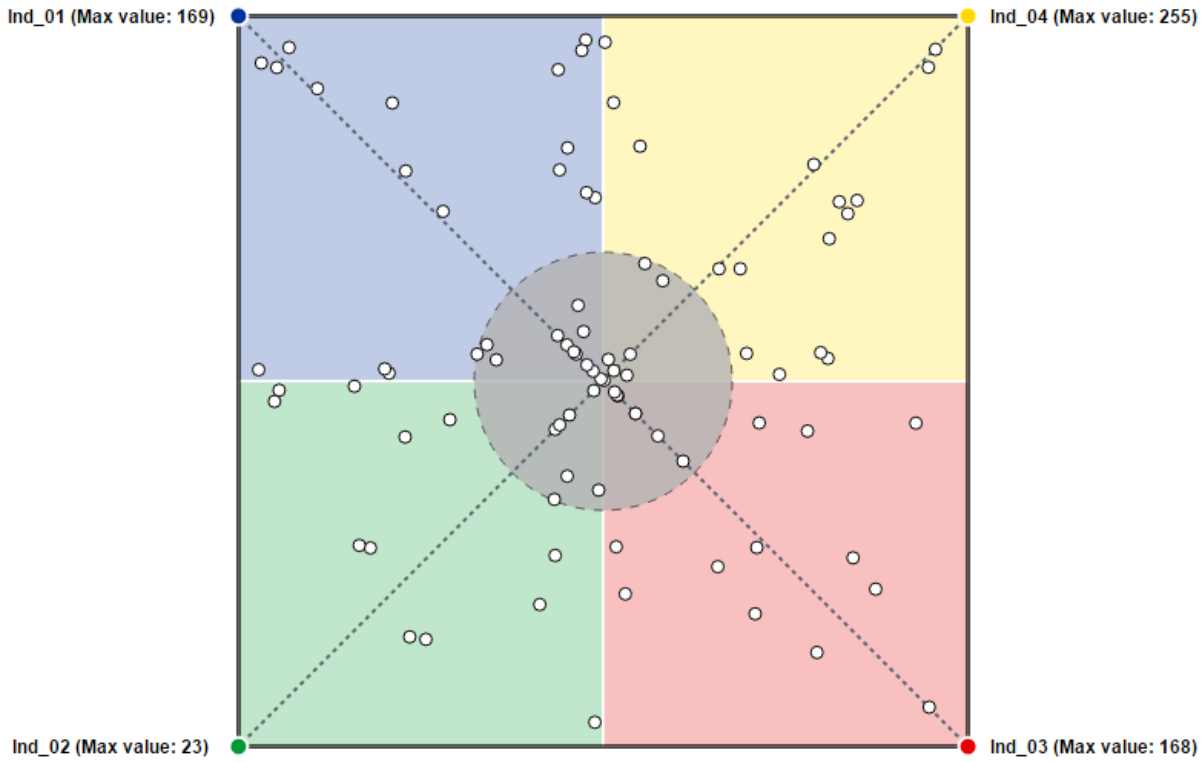

Figure 15 - A squared scatterplot: points are arranged according to the data-fit strategy.

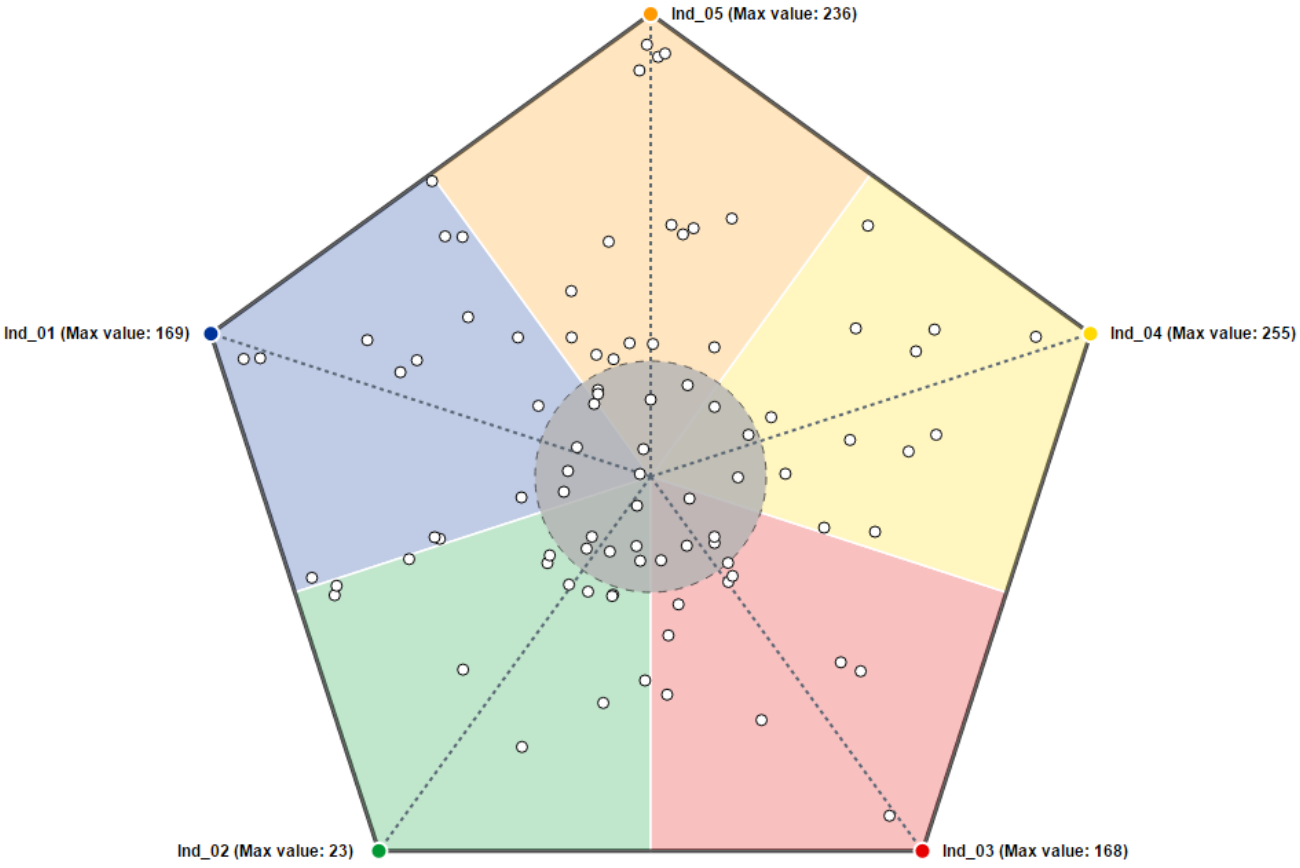

Figure 16 - A pentagonal scatterplot: points are arranged according to the data-fit strategy

Some of the factors considered to run this algorithm are:

- The number of dimensions that could have an effective attraction power (whose values are typically outside the sink threshold) for the point. For instance, if this number is zero then the point is represented in the very centre of the polygon; in case the length equals to 1, then the point is placed by the dimensions sieve approach with an alpha value set to 0.95;
- the flatness of the distribution expressed as the difference between the maximum and minimum, normalised values: in the case there is a small variance, then the weighted barycentre approach would apply (because none of the dimensions clearly outstands over the others);
- the number and quality of peaks in the distribution: all the normalised components exceeding a threshold of 0.9 are considered high peaks, while those ones exceeding a value of 0.7 are labelled as mid-peaks. For instance, in case of more than a high-peak and none mid-peaks, points are represented by the top-2 dimensions approach; while with just one peak the dimensions sieve strategy is chosen.
- This strategy allows points to occupy the polygonal space more efficiently, increasing the data-ink ratio described by Tufte in (Tufte & Graves-Morris, 1983), since they are generally placed in a sparser way according to their numeric features. In turn, since the number of loci where to focus the attention grows, it could result a little bit confusing to detect trends at a first sight. Nonetheless, it opens new possibilities in extracting new insights of the policy data.

2.3 Interactive features for points exploration

The data point representation is a crucial aspect of the polygonal scatterplot chart but all alone is not enough to provide deeper and useful insights into the data themselves. Instead, encouraging users to an active exploration of the data features is the best option to gain more knowledge from them. In this section, the main interactive features implemented to accomplish this goal are described. In particular, we omit here irrelevant and more focused issues of the interaction such as how to change the indicators to map in the chart or the point arrangement approaches, since their description is not essential to understand the chart nature and capabilities. However, these topics are described in *D6.2 – Visualisation Tools for Simulation and Scenario Analysis* on how to properly use the dashboard.

The main interactive features to highlight in this section are three and involve to play with both points and polygon vertexes. In particular, they are as follows:

- Showing tooltips about data points;
- Representing a similarity measure of the point in the polygonal space; and
- Representing the contributions of a specific dimension for each point in the polygonal space.

In the following subsections, these features are described with more details.

2.3.1 Data points tooltips

In order to give more context to the user, a tooltip listing the data features of the point hovered by the mouse pointer pops up (see **¡Error! No se encuentra el origen de la referencia.**). Such tooltip is composed by two parts, a title with the name of either the urban region or the policy alternative under exploration, and a body where the name of the actual dimensions and the corresponding, original values are presented.

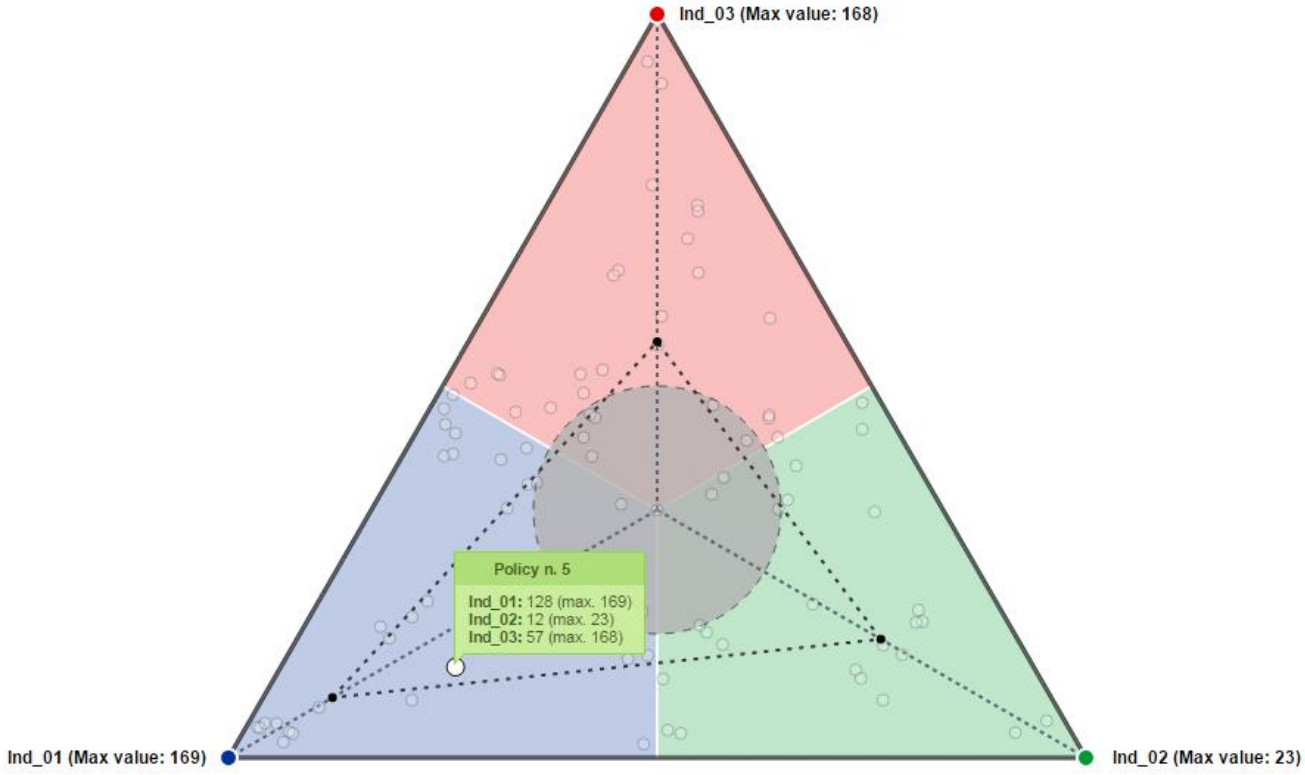

Figure 17 - A tooltip and the radar chart for the point hovered in the chart. Both these techniques represent the same information: the former in textual form, the latter visually. The reader can also note as hovering a point causes all the rest to disappear (i.e. their transparency value is set really close to zero). This is done to remove any sort of concurring attention foci that could prevent the user to focus on the selected target.

2.3.2 Similarity measure

As stated in the previous sections, one of the main interest of an analyst when dealing with multiple options of a policy is to evaluate both which dimension(s) would benefit the most across the different alternatives and if there is a subset of options leading to similar results. At the same time, to figure out how urban regions are characterised according to their dimension / objective performances, a way to compare them accurately and globally is required. The position of a point is the first feature to look at but it is not enough since adjacent pairs of points could have the corresponding dimension values far different but nonetheless could present a global, numeric behaviour very similar. This kind of situation must be avoided as far as possible because it is very likely to lead to interpretative mistakes. In other words, the principle of adjacency has to be valid just in presence of n-ples of points whose related features are really similar. Each big difference among dimensions should be explicitly represented in the chart by pushing away the points involved. In the original proposal by Chen and colleagues, the degree of similarity was not explicitly treated as a feature to present to users: it was taken into account as one of the goals to achieve, but it had not any visual validation. Perhaps, a way to derive it is to consider the compactness and the shapes of the Voronoi cells, but this is far complicated and for sure highly inaccurate. Because of both this indirect approach – that in our opinion would slow down the visual analysis of the chart – and the number of points we have to typically deal with – when representing urban areas, we could have some dozens of them, thus limiting the space where placing glyphs –, we decided to encode policy / region similarity values in a different manner. Indeed, our approach consists in:

- Showing similarity values just with respect to a chosen point;
- Encoding this value in a white-to-black colour scale that is showed to user when clicking on a specific point of the chart. The clicked point represents the touchstone all the other points are compared against and the colour assigned to it is white. The lighter the shades of the other points, the more similar to the point of reference. Black points are therefore data vectors with really different numerical features.

So, let us consider two different data vectors *v* and *w* whose components are defined as in the previous section and such that:

$$v = [v_1, v_2, \cdots, v_m],$$
$$w = [w_1, w_2, \cdots, w_m]$$

From the notation above, the similarity measures used in our platform can be described as follows:

a. **The Minkowski distance**: This measure is based on the concept of Euclidean distance but it extends to an arbitrary number of dimensions *p*. In mathematical terms, it is written as:

$$d_M = \sqrt[p]{\sum_{i=1}^{m} |v_i - w_i|^p}$$

For the purposes of our work, *p = m*. For instance, when considering three dimensions, this distance is the cubic-root of the sum of the cubic difference – in absolute value terms – of each corresponding dimension. An example of such technique is depicted in Figure 18.

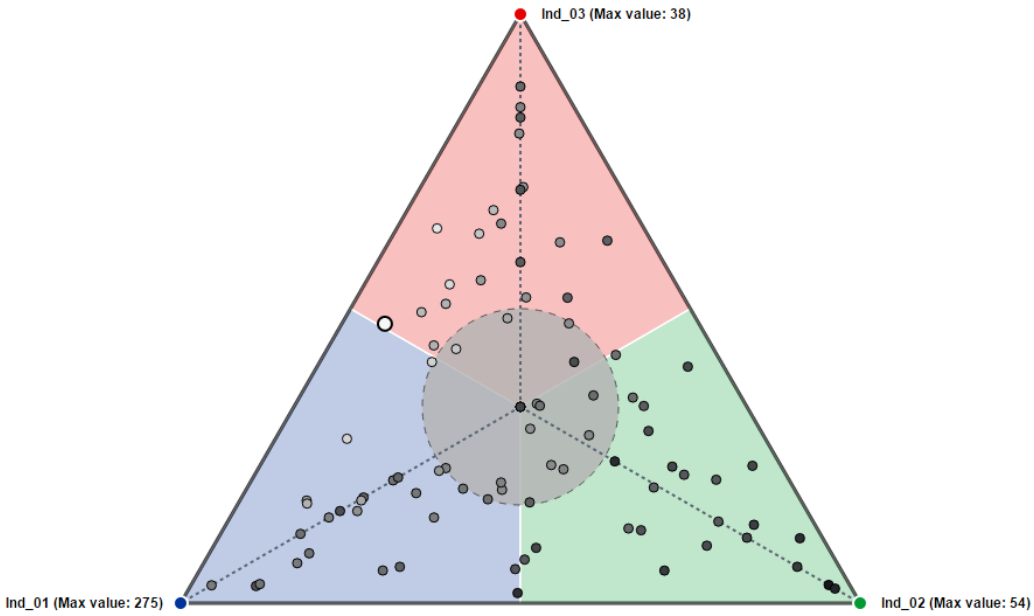

Figure 18 - Showing the Minkowski similarity between a reference vertex (the white, bigger between the red and blue areas) and the rest of the vertexes. Similar data points are placed closed to each other.

b. **The Jeffreys divergence** (Taneja, 1989): this is the symmetric version of the well-known Kullback-Leibler (KL) divergence, a measure based on the idea of Shannon entropy or information deviation (see Figure 19):

$$d_J = \sum_{i=1}^{m} (v_i - w_i) * ln \frac{v_i}{w_i}$$

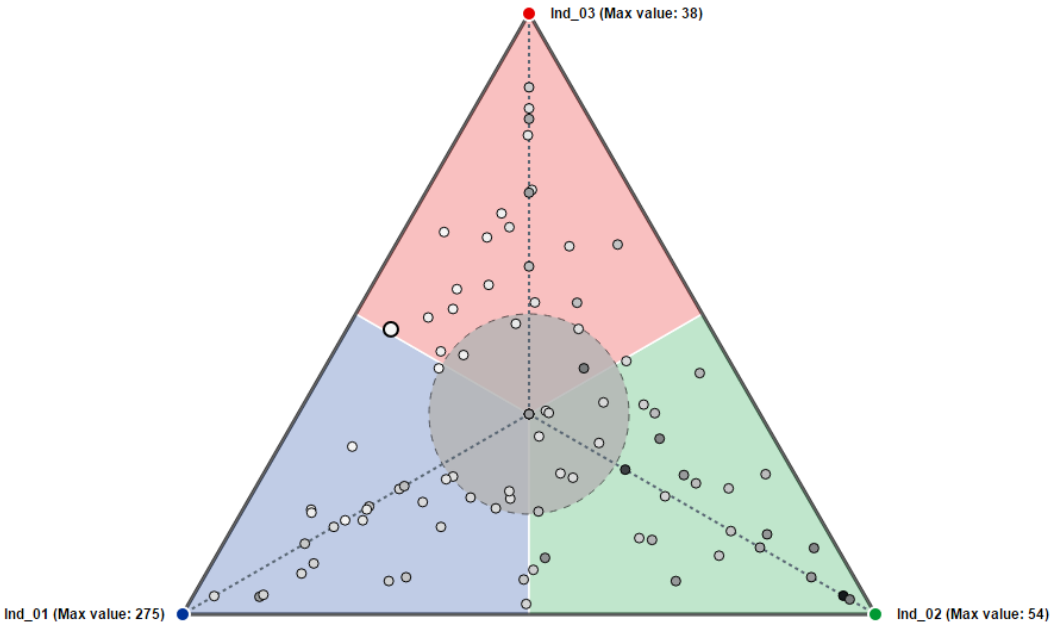

Figure 19 - Showing the Jeffrey's similarity between a reference vertex (the white, bigger one between the red and blue areas) and the rest of the vertexes. According to this metric, there are much more similar points than in the other case.

c. **The squared-chord measure** (Gavin, Oswald, Wahl, & Williams, 2003): this measure expresses the idea of fidelity similarity through the geometric means (see Figure 20). The standard form is written as follows:

$$d_{sc} = \sum_{i=1}^{m} (\sqrt{v_i} - \sqrt{w_i})^2$$

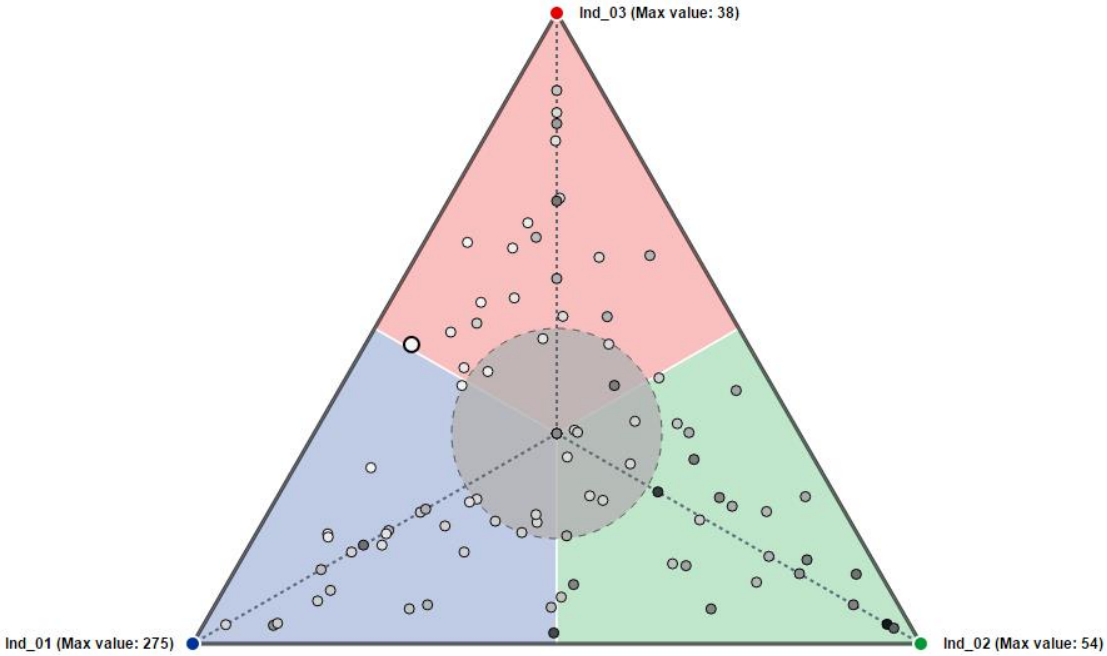

Figure 20 - Showing the squared chord similarity between a reference vertex (the white, bigger between the red and blue areas) and the rest of the vertexes. Similar results to the Jeffrey's metric are shown here. However, it appears more evident here the presence of a pair of not-so-similar points (they are a little bit overlapping in the middle of the red axis) wrongly placed side-by-side in this point distribution.

2.3.3 Contribution per dimension

Since the position of a point is determined by suitable weights for each component, one question arises: how is it possible to get an exact picture of such contributions individually? Indeed, the position provides a qualitative, though revealing, information but raw numbers are hard to derive from there. In the original proposal by Chen and colleagues, it was possible to extract this information by looking at the radial bar glyphs, where each sector was a representation of the numeric contributions per dimension. But for the same reasons outlined in the previous point, we decided to walk a different path. Indeed, by clicking on one of the polygon vertexes, all the points in the chart are coloured with a colour shade ranging from white (meaning that the point has no contributions with respect to that vertex) to the colour of the vertex itself: when these two colours correspond perfectly, it means that the point has the highest value as possible allowed for that dimension. Of course, shades in the between proportionally depict different contributions with respect to the maximum value (see Figure 21, Figure 22 and Figure 23).

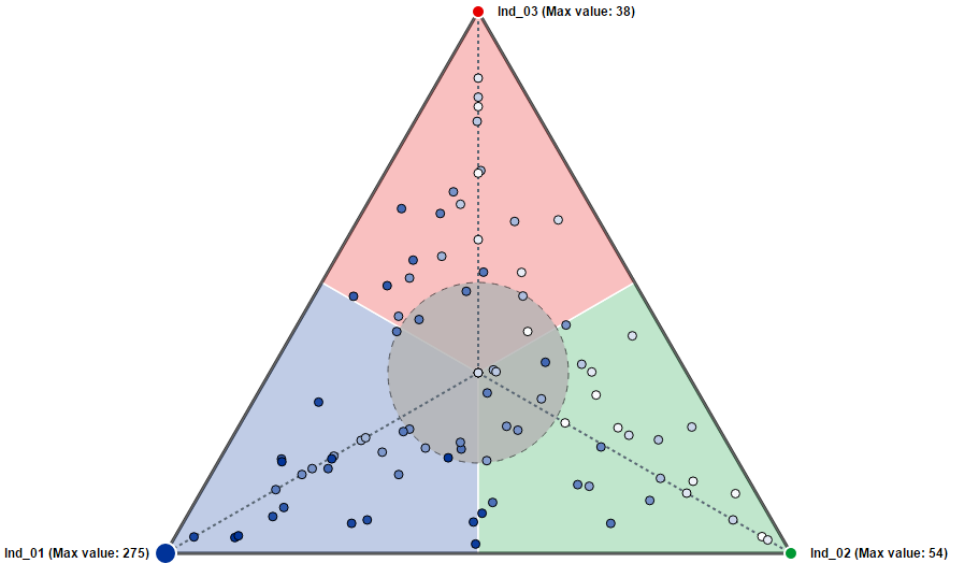

Figure 21 - Showing the values of dimension 'Ind_01' for each point of the chart: the points with higher values in that dimension are in average much closer to the vertex of reference. The point arrangement strategy here is weighted barycentre

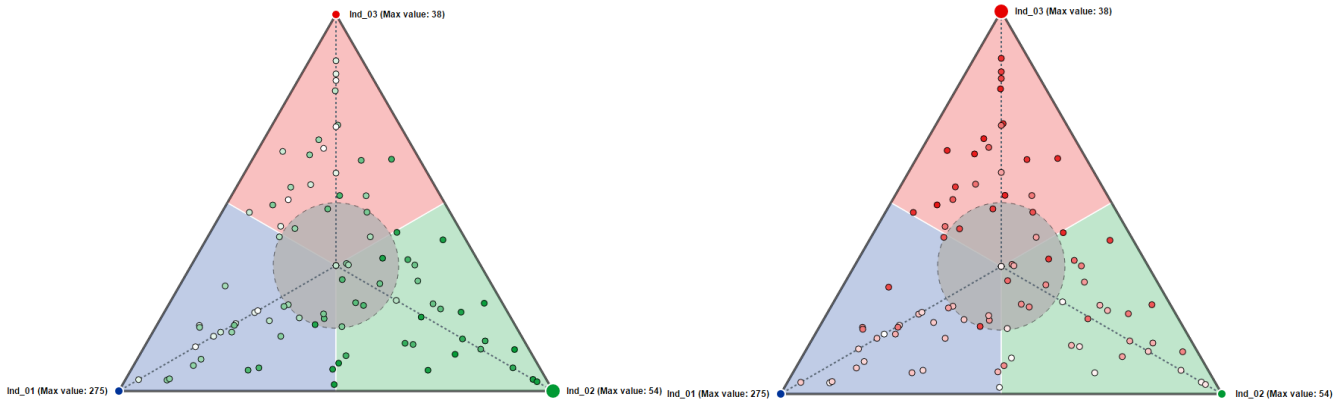

Figure 22 - The same as in Figure 21 but for the other two dimensions: 'Ind_02' on the left and 'Ind_03' on the right. It seems that some points could be classified as outliers since they are placed too further from the vertex of reference with respect to the value they have in that dimension. Instead, it means that the other two dimensions have a better contribution each for those points.

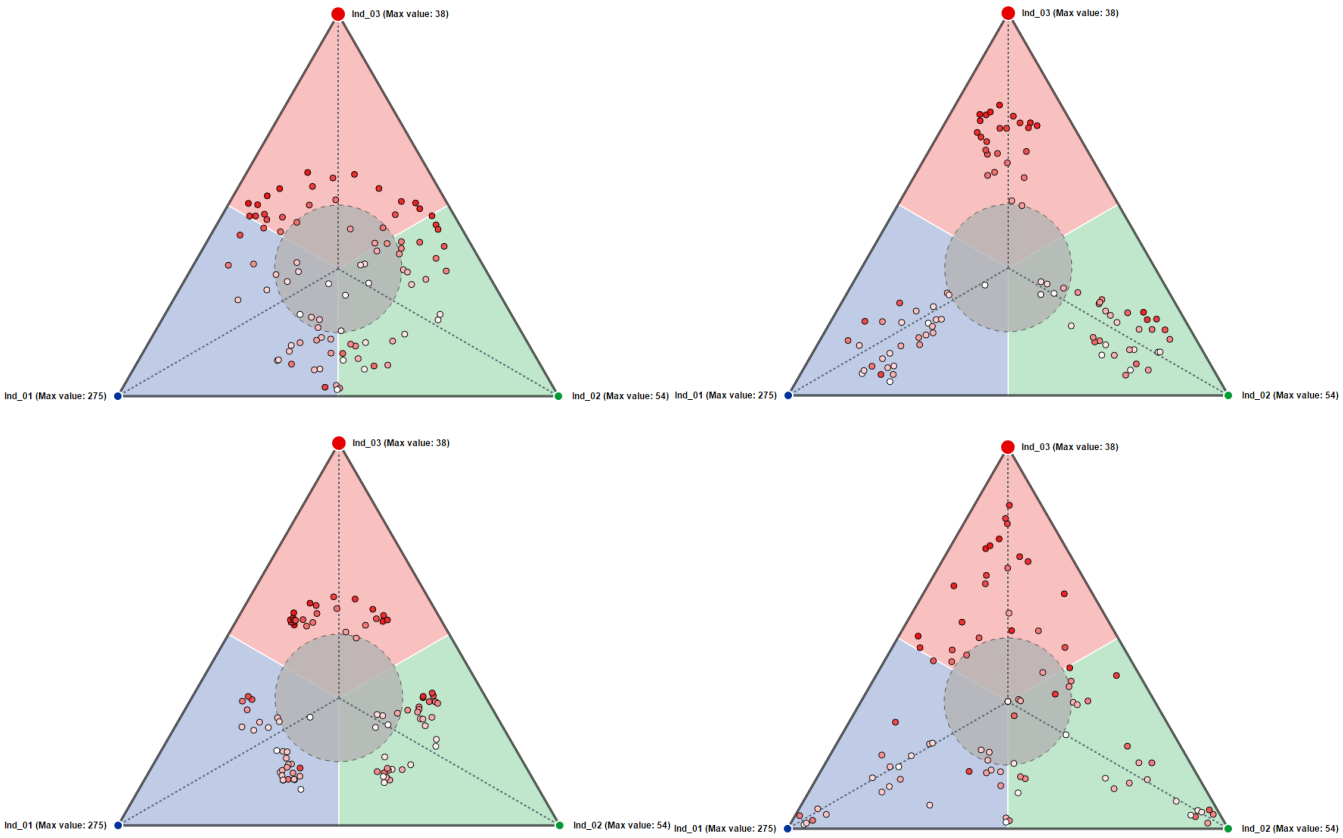

Figure 23 - Analysing contributions of a dimension across different point location strategies: from top to bottom and left to right, top-2 dimensions, dimensions sieve, similarity enhancement, and data-fit approaches are presented. In general, the orientation property firmly holds for each representation chosen. Some unexpected results would suggest further exploration about the relationships with other dimensions.

To reinforce this kind of discovery and allow at the same time to make comparisons across the dimensions, another feature has been implemented. When hovering a point with the mouse pointer, it is possible to see the radar chart corresponding to that point such that the values for each axis will be highlighted (see Figure 24). This way, users can get a more detailed idea about the different contributions, the possible trade-offs and the value distribution of each of the points within the polygonal space.

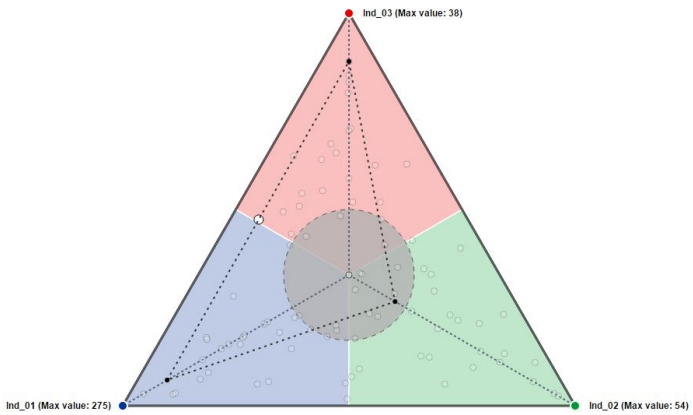

Figure 24 - By hovering a point, it is possible to see its whole contributions for each dimension. In this case, it is evident that the blue and red values are really high while the green one is not passing the sink threshold. According to this, it is normal to see that point close in between the blue and red Voronoi cells.

2.4 Advantages and drawbacks

The chart described throughout this chapter was introduced as a way to represent multi-dimensional data and provide the tools to reason about trade-offs among policies alternatives aiming at satisfying conflicting objectives as well as to characterise geographical regions with respect to different points of view. Table 1 and Table 2 summarise the main commonalities and differences of our proposal with respect to the one by Chen et al. (2013).

Table 1 - Summary of the main similarities in the approaches by (Chen et al., 2013) and our proposal (see text for details)

Similarity	Explanation
Polygonal layout	The chart has a regular polygonal shape where each vertex represents a dimension of the policy. The polygon sides mark the space available to draw the points.
Emphasis on orientation and adjacency properties	Points are arranged into the polygonal space trying to maximise both *orientation* (i.e. the highest the contribution of a dimension, the closest the point to the corresponding vertex) and *adjacency* properties (i.e. similar policies should be placed the closest as possible to each other.
Presenting solution by *a posteriori* approach	All the possible policy alternatives are given to the user at the beginning. Filtering and searching operations reduces the search space during the analysis / exploration task.
Reducing dimensionality	Policies are depicted according to a subset of m out of n ($m \leq n$) dimensions.
Allowing partial order of the policy alternatives	From a mathematical point of view, vectors with multiple dimensions cannot be totally ordered and as a consequence, the visualisation technique has to reflect this property. In particular, our proposal can interactively display the contribution of each alternative given a selected dimension while keeping the picture of the broader context under analysis.

Table 2 - Summary of the main differences in the approaches by (Chen et al., 2013) and our proposal (see text for more details)

Difference	Explanation
No Self-Organizing Maps (SOMs) approach to place points	SOM approaches to automatically learn the optimal position where to place policy points based on their dimension values, have been extensively used in a number of visualisation approaches. However, they introduce one more layer of complexity to the overall technique. To keep our implementation simple and more flexible, we rely on a set of lighter algorithms that could be used alternatively to highlight different data properties.
No *a-priori* vertex arrangement	Vertexes are not optimally arranged to minimise the topological distortion of the point arrangements within polygons having more than 4 sides. Instead, users can choose the assignment vertex-dimension as they wish. This approach simplifies the algorithmic complexity of the implementation but it could lead to some visual issues (e.g. inconsistencies, biased pattern reveals) according to the permutation chosen.

Polygonal inner space is not organised through Voronoi cells based on policy alternatives clusters	Voronoi polygons partitioning the inner space of the chart determine the 'influence areas' of each dimension and therefore they are computed from each polygon vertex.
Explicit sink area drawn	The sink, namely that area in the chart where well-balanced alternatives are likely to fall regardless of the raw values of each dimension, is explicitly shown.
No glyphs associated to points to resume policy dimension contributions	For a sake of compactness and clean presentation, policy alternatives are displayed as small, white points instead of polar area charts. This is particularly useful when several alternatives are presented at the glance. To have an overall view of the contribution of each dimension to a single policy, it is possible to hover each point with the mouse pointer to see the corresponding radar chart.
Any prior computation of the Pareto Frontier	We show all the policies in the chart without a previous step of computing the Pareto Frontier to make the visualisation platform faster and lighter
Polygonal scatterplot with a double usage	The original proposal intended only to characterise and evaluate policies across several dimensions / objectives. We extend this use to compare how the different zone of a city react to different policy objectives

To support the analysis in an effective manner, each chart must be applied in the correct context, on the right data and to solve a (limited) number of specific questions. Considering this principle, this section illustrates some of the pros and cons related to the use of our version of the polygonal scatterplot (Table 3 and Table 4 respectively). The discussion is based on two main sources; our direct experience during the implementation and testing phases and the feedback provided by INSIGHT partners and collaborators in response to specific showcases or after manipulating the dashboard by themselves. A formal investigation on the user experience with the tool will be launched in the near future to gain more insight into the actual use and applicability of the chart.

Table 3 - Description of the main advantages to introduce polygonal scatterplot charts

Advantages	Description
Explicit representation of trade-offs	The point arrangement strategies have been studied to maximise the perception of the different compromises among the dimensions to take into account. For this end, the metaphor of 'gravitational laws' exerting on each point is particularly useful to describe the grounding idea about the chart.
Good balance of qualitative and quantitative analysis	Addressing properties such as orientation and adjacency when conceiving the point placement approaches help users to spot the nature of the trade-offs as well as to find a global characterisation of the data points according to some dimensions. This way, comparisons and the search of patterns could be supported. The presence of tooltips and vertexes labels add some numerical insight to the qualitative discoveries.

Different strategies to arrange points leads to different insights in data patterns	Multi-dimensional data are hard to treat because of the number of dimensions (and their possible 2^n combinations, if n is the original number of dimensions) they have. To keep the complexity as low as possible and maximise the capability of visual reasoning, we have proposed a chart where it is possible to: • easily explore several combinations of m dimensions (at least for values of m such that $m \leq 5 \leq n$); • change the placement laws in order to highlight different aspects of the data patterns and have a more complete view of the data domain under analysis
Doubling a typical scatterplot capacity	By definition, scatterplots are bi-dimensional since they are represented in a (x,y)-Cartesian space. By changing some of the construction rules, we expanded their capacity to deal with more than 2 variables. By the way, it is still possible to represent just two variables according to our approach (or even just one). In this two cases, the basic shape would be a triangle as well and points will be drawn in just two (or one) Voronoi cells.
Compact representation	Our proposal has a compact representation, even better than a traditional radar chart. Indeed, in the latter case all the circular space of radius r is used to represent the data, while in our chart we limit ourselves to the space inside the polygon (which is in turn externally delimited by the same circle of radius r where the regular figures are inscribed into). Moreover, adding a dimension (polygon side) would increase only the total area of the polygon at our disposal without affecting the external circular area where the polygon lays in. This is different from parallel coordinate charts since adding a dimension there means extending the space to represent data (typically by adding a new vertical axis)

Table 4 - Description of the main drawbacks to introduce polygonal scatterplot charts

Drawbacks	Description
Some initial difficulty in correctly interpreting the chart	Being a relatively new proposal (at the best of our knowledge there are no other works dealing with this approach apart from those of Chen and colleagues), it is natural to expect that some interpretative difficulty arises. Moreover, since scatterplots are a well-established data visualisation technique, the structural changes introduced here could be especially difficult to digest. Thus, changes in the mind-set that are required to accept and correctly interpret this new chart could need some time. In this sense, further studies on perceptive issues and user experience with this representation technique would be very useful to inform future developments and improvements.
Well-balanced points are likely to be mapped in the centre of the sink, regardless of the variable magnitudes	Filters being able to show only the best data options (for instance those whose components are all passing the sink threshold) could help to mitigate such inherent issue of the chart. Moreover, showing point similarity values as well as hovering points to explore their radar chart shape are two other helpful tools to correctly interpret the result mapped in the chart.

The vertex arrangement could be critical for *n>4* and more than 5 or 6 dimensions could mislead the interpretation of the chart	When points are subject to some forces exerted by vertexes placed at the very far ends of the chart, their placements could lead to incorrect evidence, as already noted in (Chen et al., 2013). In these cases, the properties of adjacency and orientation do not hold anymore. All the explorative techniques we implemented and described above go to the direction of mitigating negative effects on user's reasoning. At the same time, the different strategies to place points have been introduced to overcome the issue. Further efforts shall be devoted to find the best vertex arrangement(s) that minimises the number of opposite forces. In the current version, the vertex disposition can be changed manually.
Difficulty in detecting relationships between pairs of variables as in the traditional scatterplot	Scatterplots are used to detect if two variables have some kind of 'simple' relationship (e.g. (sub-) linear, exponential, logarithmic) among them. Our version of scatterplots loses the most of this capability since the points coordinates are computed as a function of the original, *m*-dimensional values and they are not directly determined by the values themselves (and axes are not orthogonal to each other). In any case, this is not the principle according to we conceived the chart and therefore it cannot be used as such.

3. Using the visualization platform to support policy making: cordon tolls in Madrid

In this section we showcase an example, using the outputs of the cordon toll policy for Madrid, to illustrate how the INSIGHT visualisation platform can be used as a policy decision support tool. We focus on demonstrating how the different tools put in place by the platform can be used in a linked and complementary manner to compare alternative policy measures and derive insights from the corresponding simulation outputs. Thus, rather than on drawing conclusions about the data depicted, the emphasis is on highlighting the main functionalities of the dashboard and the different flavours of the charts when the platform is brought into operational use.

For this example, let us consider the total number of jobs predicted by the simulator for each zone of the Metropolitan Region of Madrid, and let see its evolution across time (see Figure 25 for a visual reference). We recall here that the territory of Madrid under analysis has been split into a set of 90 smaller regions of interest according to the following criteria: i) The city of Madrid is divided into the original 21 districts; ii) The remaining zones could represent a single town, an agglomeration of adjacent towns / small villages, or part of a bigger city. For more details, please refer to Section 4.1 in *D6.1 – Visual Ecosystem Technical Specification and Design*.

In this section we analyse and compare two policy scenarios. These correspond to the baseline condition for the city of Madrid and to the one resulting from placing a cordon toll around the city, all along the M40 beltway. As a reminder, the baseline scenario represents how would evolve the present situation without any policy intervention, whereas the M40 cordon toll scenario describes what would happen (according to simulations) when the M40 cordon toll policy is put in place (i.e. users have to pay a fee every time they cross the M40 beltway).

With this frame in mind, let us get back to Figure 25. The first evidence standing out is that the two charts are pretty similar and indeed any noticeable variation seems to happen during the years portrayed in the figures. A misleading conclusion would be that the introduction of a cordon toll at M40 does not have any impact on employment. But, when looking at the numbers along the vertical axis, it is possible to see that the chart depicts values greater than 160000 jobs per region, which it means that small variations in the job count cannot be observed at a first sight. Another interesting point is to see that just a small fraction of the regions has a relevant increase foreseen in employment, especially in the period from 2015 to 2020, which represents a first check point to evaluate the effects of a policy on a territory. Last, most of the regions considered in the study do not exceed 20000 job units and, at least apparently, none of them seems to be affected by the job growth spotted before.

By applying a spatial filter, it is possible to analyse the data for those regions laying in specific areas of interest. For instance, in Figure 26, we decided to show just the 21 districts forming the city of Madrid and highlight those having the highest number of employments. It is possible to see how the city is divided into at least two parts, if considering the job distribution: indeed, the southern part is the one experiencing the greater problems in terms of number of jobs, while in the north and centre the situation is much better, with some remarkable contributions of those three zones depicted in the darkest orange shade in the map and corresponding to the districts of Centro, Salamanca, and Chamartin.

By looking at the map legend, it is also possible to compute the job spread between the top-3 and last-3 regions, that is the ratio of jobs presents for each pair of zones. It results that this value ranges from 7 to about 32, which

suggests to the analyst that the best district has a number of employments which is more than 30 times greater than in the worst one.

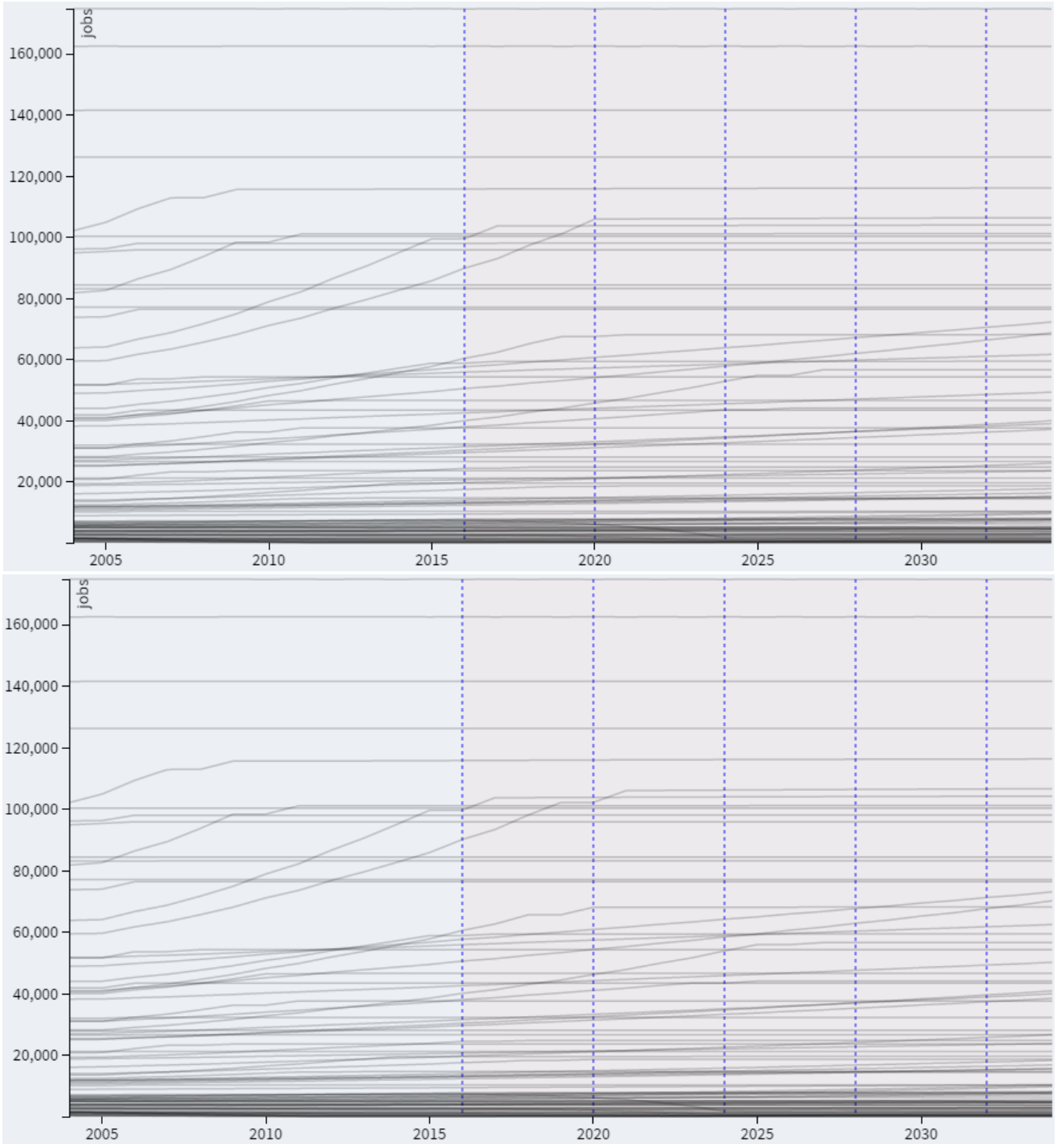

Figure 25 - Evolution over 30 years of the number of jobs per zone. Each line represents a different region in the Metropolitan Area of Madrid. From top to bottom, the graphs show the evolution predicted for the baseline and M40_toll scenario.

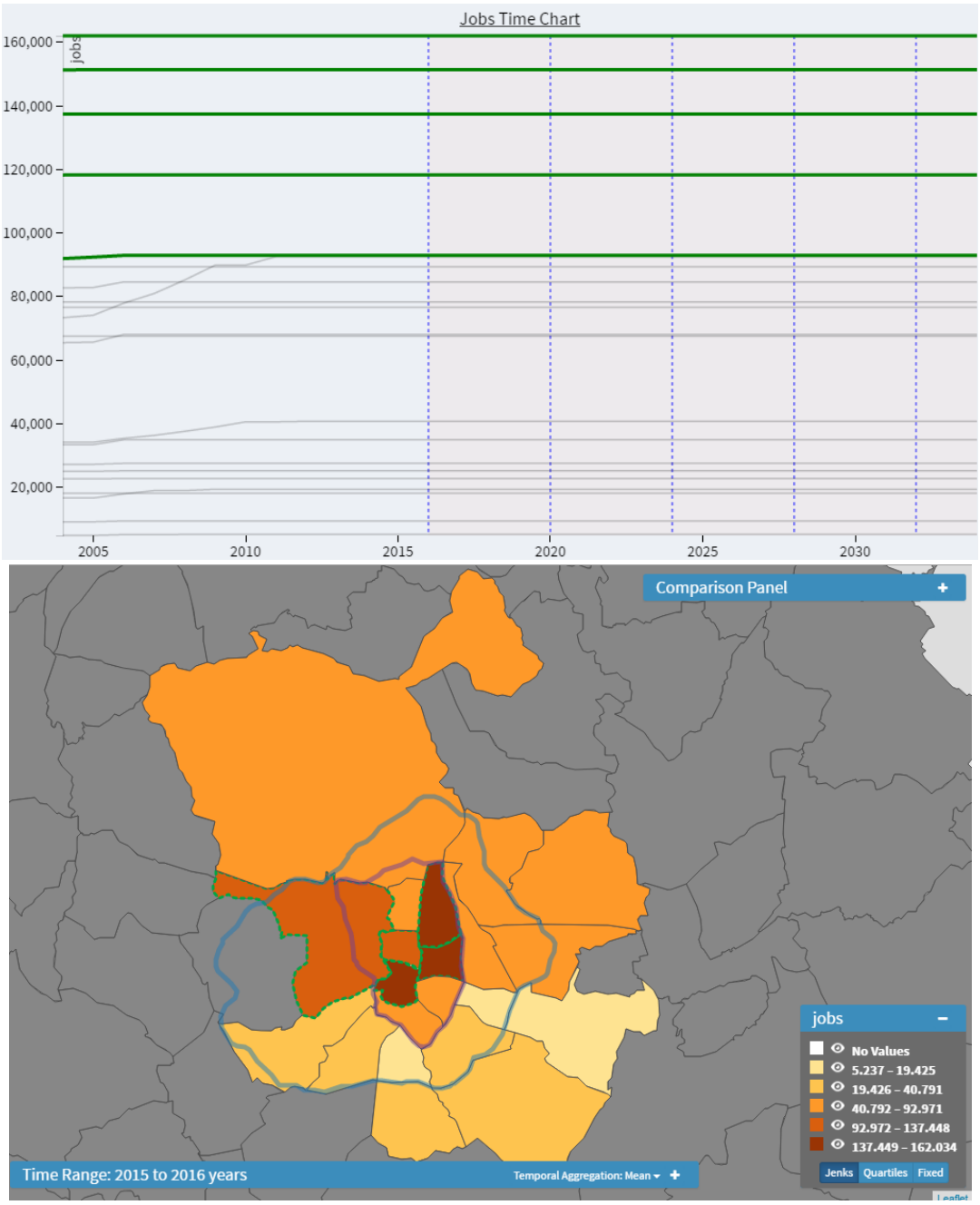

Figure 26 - Details on number of jobs in the districts of Madrid for the M40 cordon toll alternative. Above, the top 4 districts with more jobs and, below, their map representation (values for the year 2016). The districts under examination are highlited with a green, dotted line in the map.

A similar trend could be also detected by considering the evolution of the job indicator in the zones outside the city core, as shown from Figure 27 to Figure 30. In these cases, we split the regions according to their geographical position relative to Madrid, so that we can analyse zones in the North, East, South, and West of the Metropolitan Region. In the four cases it is possible to identify in the map those zones with the greatest estimated number of jobs for a given year and also, those that are expected to be less affected by the M40 policy measure. In particular, it seems that the regions taking the greatest advantage are those located in the periphery of Madrid (in a darker shade of orange in the map): this evidence suggests that the introduction of a toll to enter in Madrid would foster

the creation of new business spots (or the displacement of existing ones) in/to zones where the toll has not to be paid, but close enough to the city itself in order to take advantage of the whole city infrastructures.

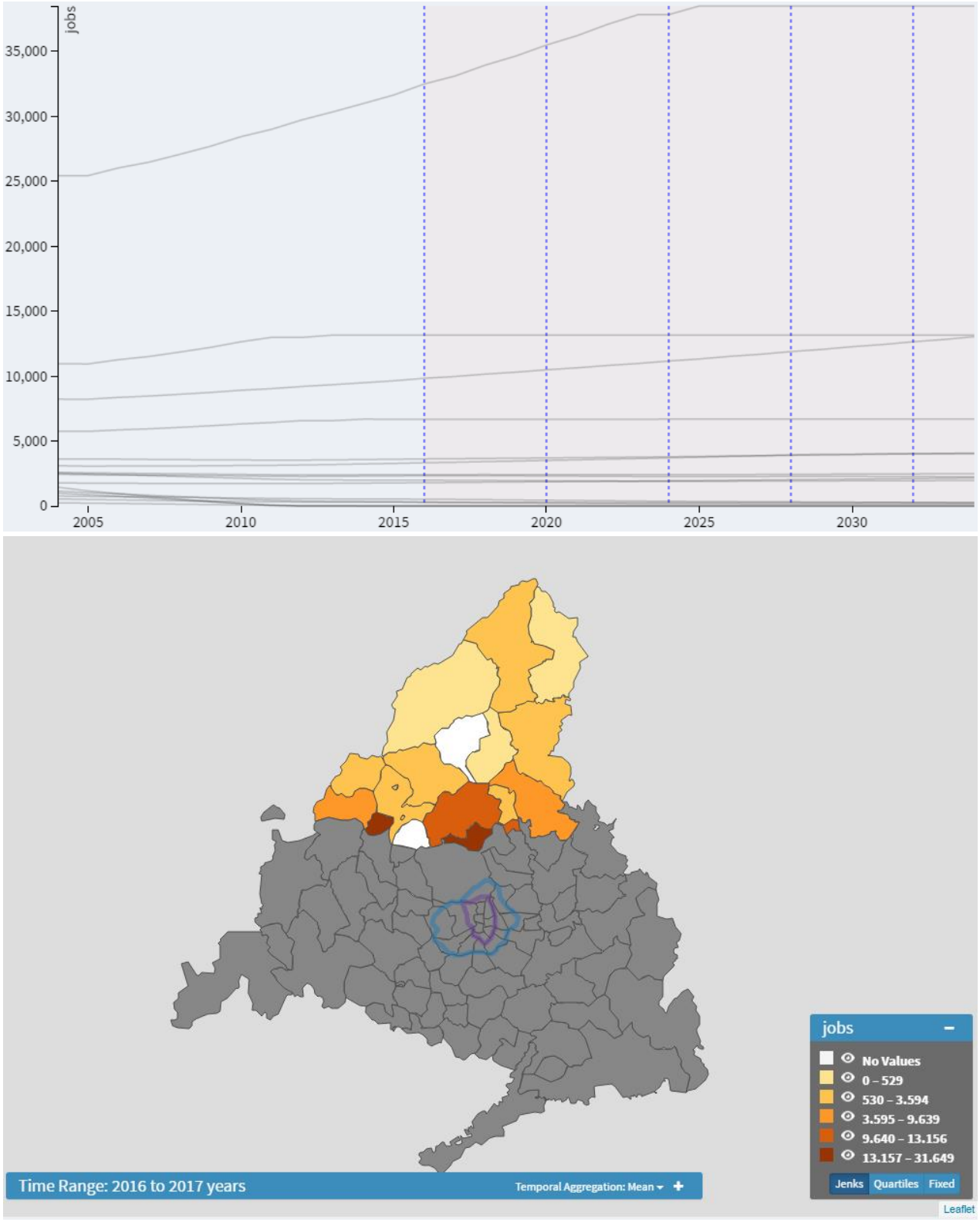

Figure 27 - Number of jobs evolution for areas in the northern Metropolitan Region of Madrid: above, time line chart and, below, values for the year 2016. All the values shown from this Figure to Figure 30 refer to the M40 cordon toll policy.

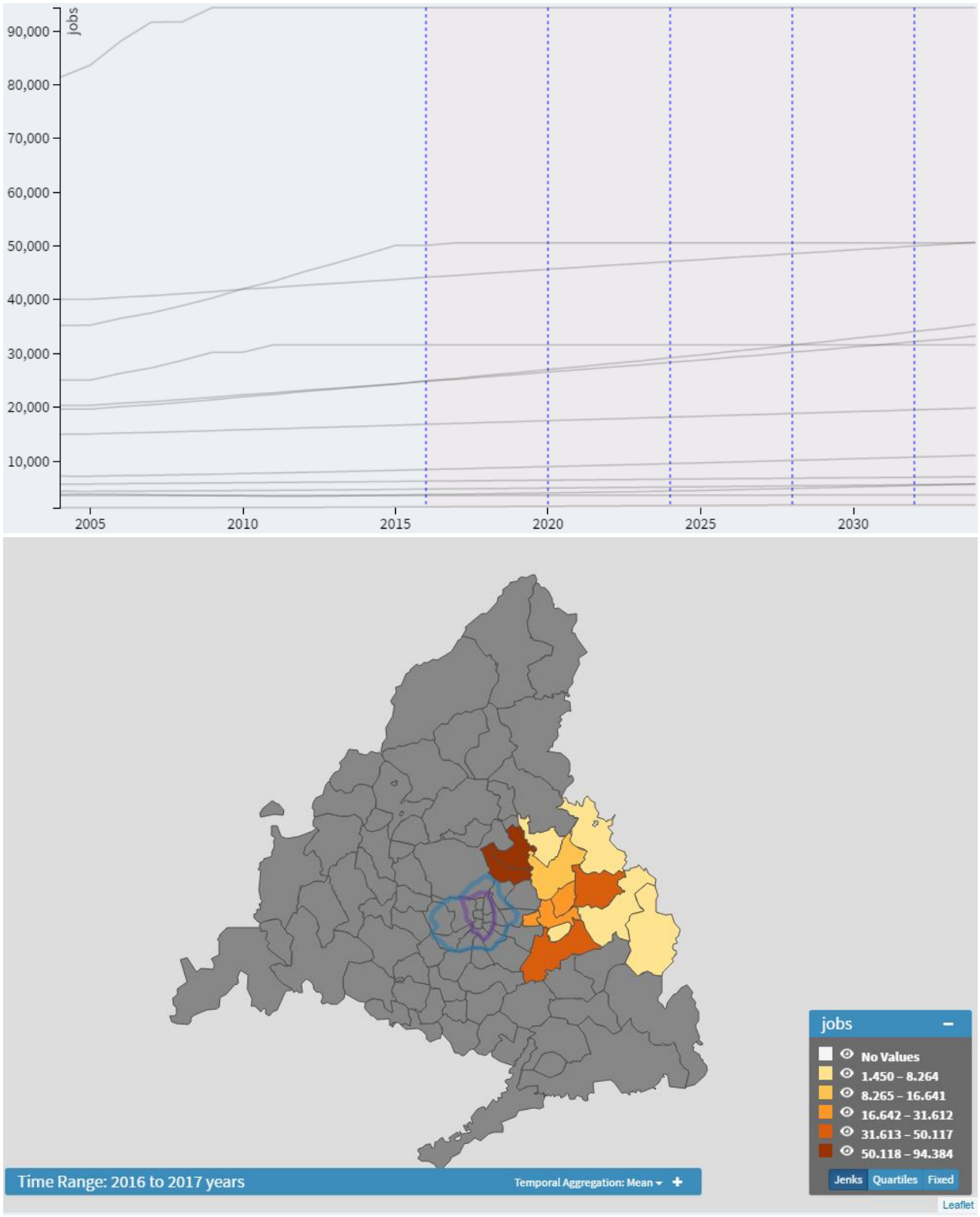

Figure 28 - Number of jobs evolution for areas in the eastern Metropolitan Region of Madrid: above, time line chart and, below, values for the year 2016.

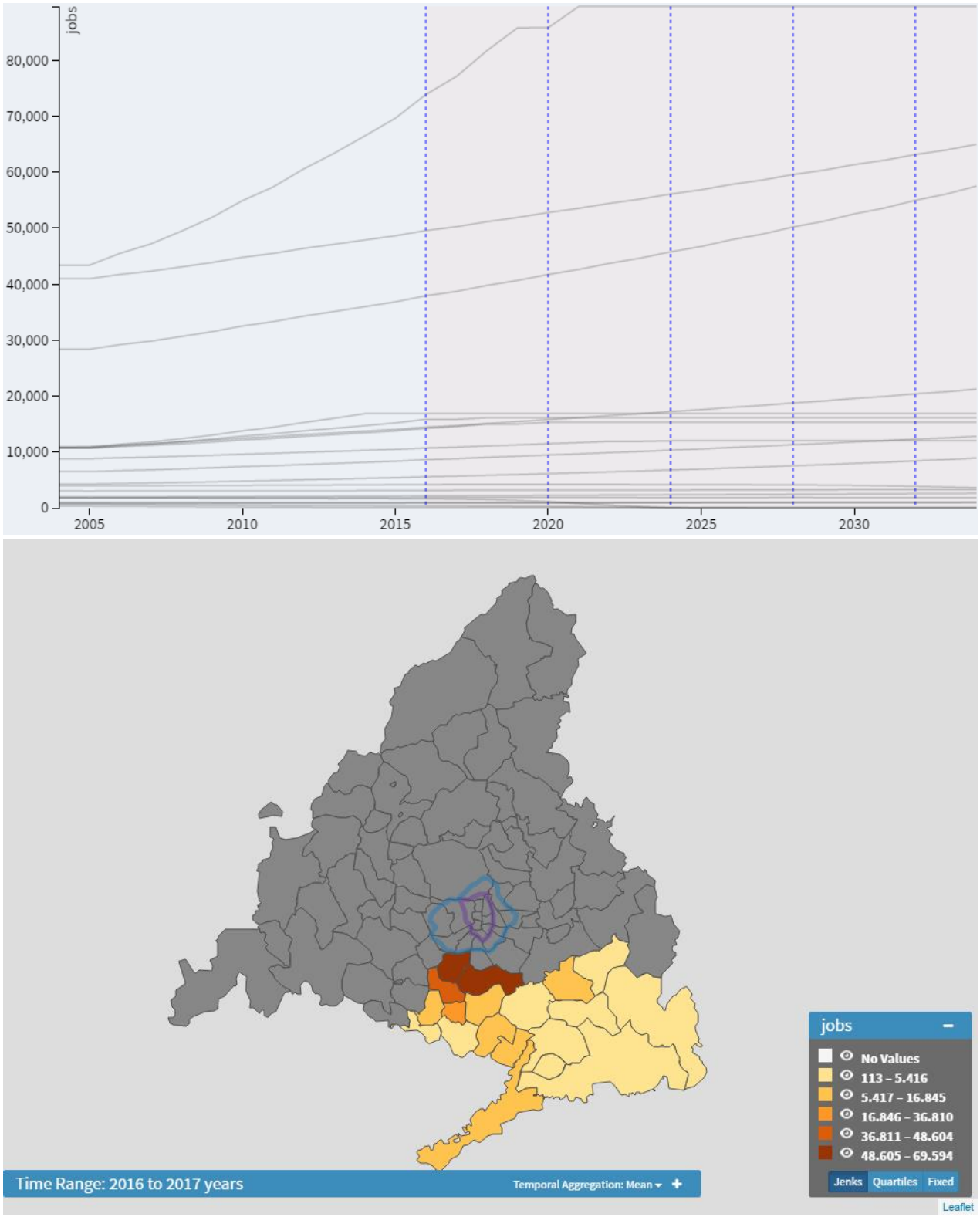

Figure 29 - Number of jobs evolution for areas in the southern Metropolitan Region of Madrid: above, time line chart and, below, values for the year 2016.

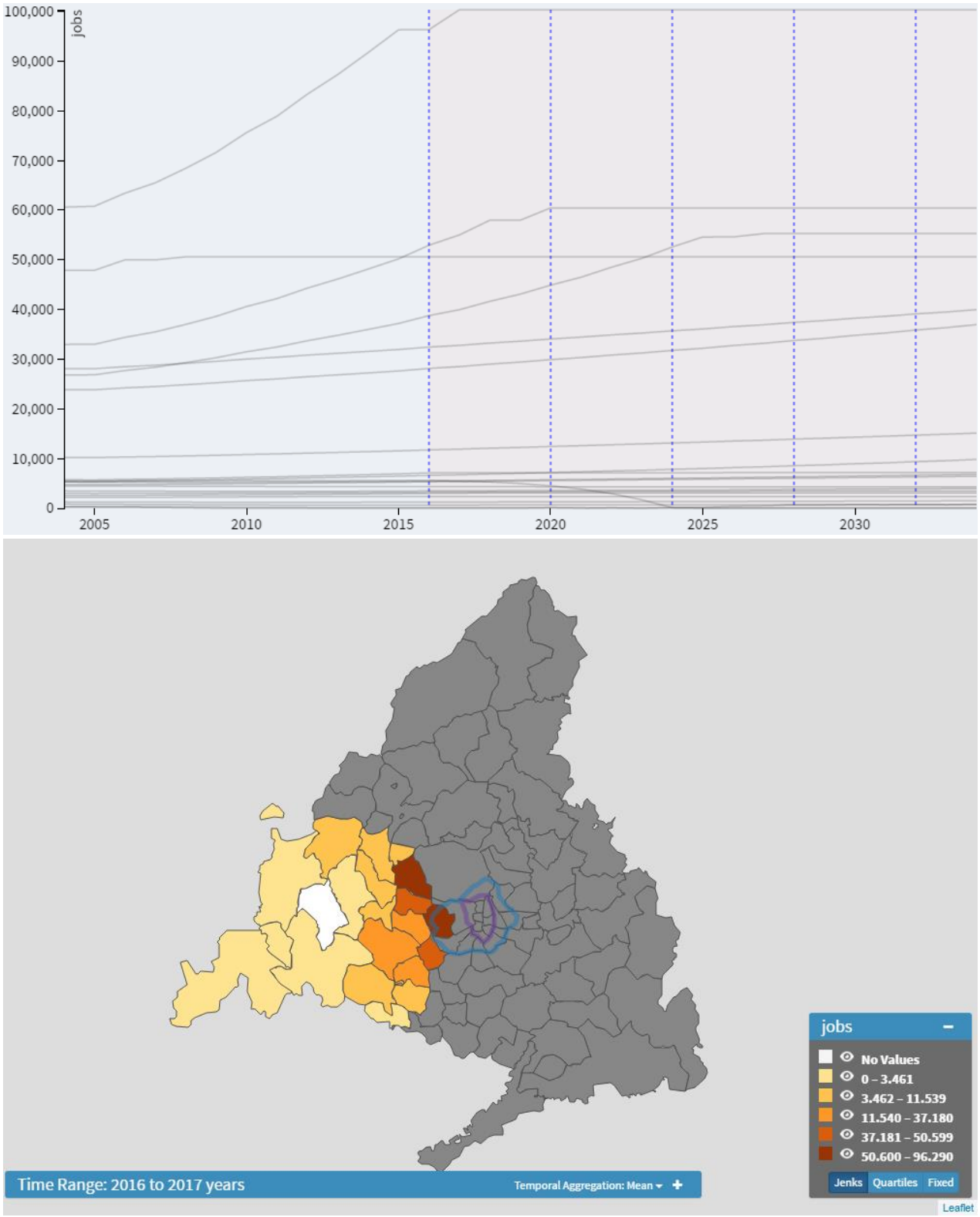

Figure 30 - Number of jobs evolution for areas in the western Metropolitan Region of Madrid: above, time line chart and, below, values for the year 2016

We can derive similar conclusions by looking at the choropleth map displayed in Figure 31 where the whole set of regions is shown at the same time. By comparing the latter representation with the ones in the previous figures,

we can note that the same region could be coloured with different shades of orange, even if the data represented in there are the same and considering that no other changes have been done. This fact could seem a visual inconsistency, but it actually is in agreement with the functionality of the filters described in *D6.2*. In particular, when filtering out some regions, the charts displays a data representation which is consistent with the subset of data left under analysis. In the case of the map, this leads to an arrangement of the statistical intervals which, at its turn, causes an adjustment in their depiction. Furthermore, the same happens with the line chart: in this case, the lines corresponding to the hidden data are removed and the scale of the vertical axis is changed accordingly to reflect the new situation. In both the aforementioned cases, it is possible to perform a double kind of analysis: on the one hand, by considering the overall picture, the contribution of each part is compared against / measured over the whole and, on the other, the filters allow to reveal local / smaller trends or patterns that could be hidden otherwise. For instance, with respect to Figure 31, it is clear where the greater number of jobs could be found, that is in the central districts of Madrid and in the region corresponding to Alcobendas at the North-East of Madrid. Since the detected areas in the capital city show a mix of great concentration of touristic places and business centres (especially big brands and (multi-)national corporations), the charts suggest to the analyst that the introduction of a cordon toll at M40 level does not affect too much the number of jobs related to tourism and services, as it could also be noted by looking at the line chart in Figure 26, where the 30-year evolution of such indicator is almost constant.

Another interesting insight could be derived by considering the line charts presented from Figure 27 to Figure 30, where the evolution of the job indicator is shown according to different areas of the region of study. If the problem in Figure 25 was that it was almost impossible to see trend details at a finer level, the above mentioned images help to clarify this point. In particular, it is evident how interesting growing patterns pop up to our attention affecting a bunch of regions per area considered. As a side note, these are the same regions identified by looking just at the corresponding maps. This fact shows one more time the importance of a fundamental concept in which our visualisation dashboard relies on: to provide different views of the same data to reveal possible patterns hidden across their multiple dimensions. The line charts under consideration could help the analyst to examine further details, which might help to characterise those zones experiencing the highest rates of jobs growth. To estimate it, a visual proxy would be to pay attention, for each line, to its slope length and direction: the longer the slope, the more long-lasting the effect; and if the slope rises towards the rightmost side, then here it is a positive trend, otherwise, this would be defined as a decrease or negative effect. Further information, such as the values on the vertical axis as well as tooltips will further add some more contextual information. According to that, the towns of Tres Cantos, Alcobendas (at least the northern part), Getafe, and Pozuelo de Alarcón (respectively in the North, East, South, and West of Madrid) show the greatest increments in their areas. Very few negative trends are spotted there: the most evident is depicted in Figure 30 in the period from 2020 to 2025 corresponding to the towns of Arroyomolinos and Moraleja de Enmedio.

Another small decrease can be found for some regions in the northern part of the Metropolitan Region (see Figure 27) from the beginning of the period of reference till about 2010. According to these charts, the introduction of a toll would benefit a number of regions in terms of number of jobs and, above all, the positive trend would continue for several years (with less evident effects as we extend the foreseen period progressively). The downside, instead, would be represented by an accentuation of the actual spread between zones in terms of job opportunities, which at its turn, could lead to some problems related to the socio-economical background (e.g poverty and segregated areas).

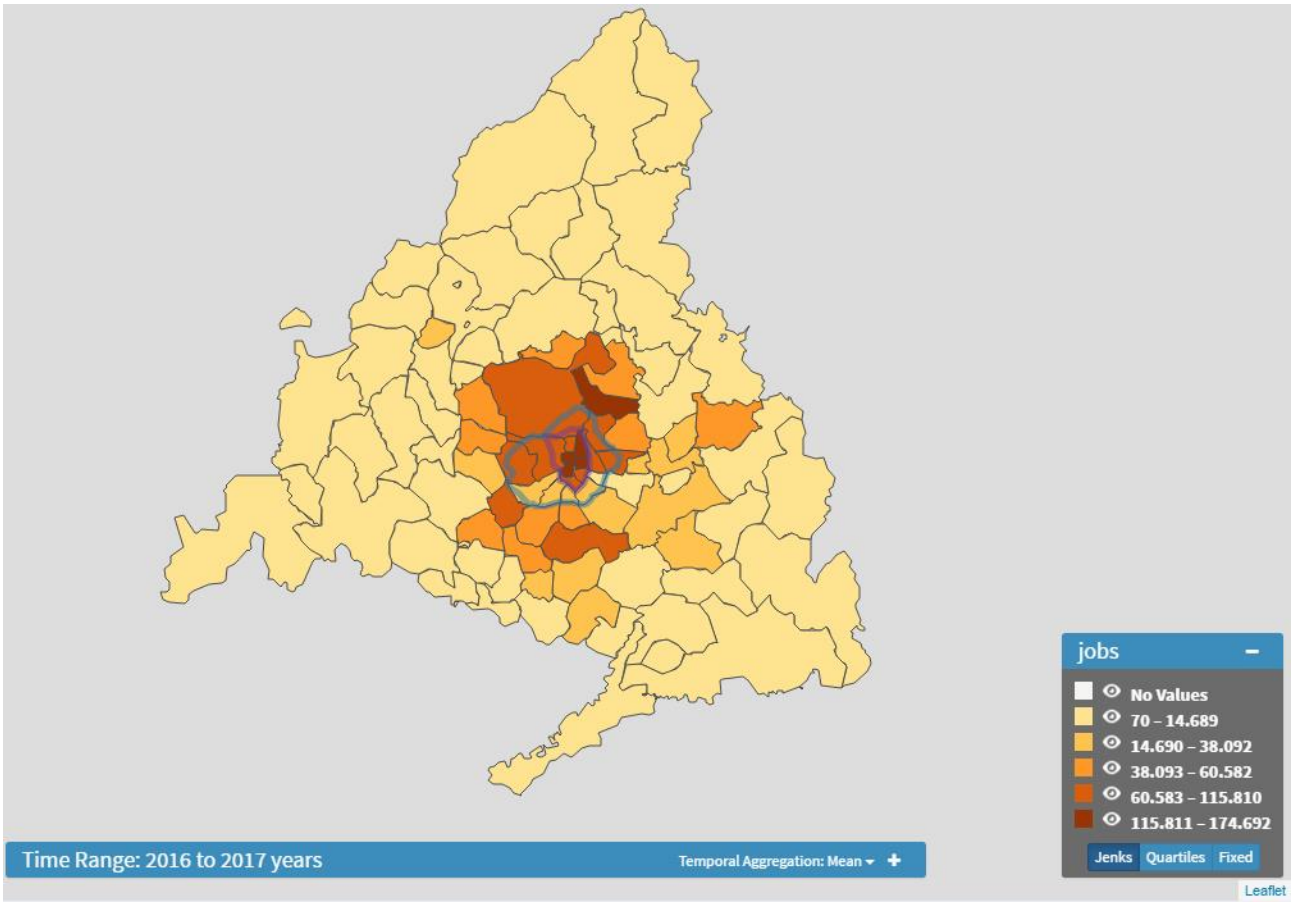

Figure 31 – Number of jobs per zone in the area of study resulting from the application of the M40 cordon toll policy in Madrid, as estimated by MARS for the year 2016. In this choropleth map, it is possible to compare the employment degrees across the regional fabric. As shown, there are zones and towns with a high concentration of jobs both in the metropolitan area (especially in the northern and central districts) and in the eastern and southern part of the periphery, respectively. Even if the colors assigned to each region change a little bit across the maps from Figure 26 to Figure 31, the conclusions drawn here and in the text are perfectly consistent among each other.

To better appreciate the differences among trends, the line chart could be used to map the alterations occurring per zones and with respect to a given year of reference, as shown from Figure 32 to Figure 34. The rationale of this representation is to change the point of view previously described to consider the effects of a policy compared to a specific point in time. This way, it is possible to evaluate how good / poor a policy alternative is to solve a problem at different temporal moments after its application. In mathematical terms, this chart shows the results of the following function:

$$\Delta_{jobs} = jobs_{y}^{z_i} - jobs_{Yr}^{z_i},$$

where *Yr* is the year of reference (in the figures aforementioned, this value is 2016), *y* is the year of comparison and z_i is the *i*-th zone to consider. According to the sign of the difference and the position of the year *y* in the temporal line with respect to the year of reference, four possible interpretations arise, namely:

- $\Delta_{jobs} > 0, y < Yr$: more jobs in the past than in the year of reference: it implies that the policy has a bad impact on the situation under analysis;
- $\Delta_{jobs} > 0, y > Yr$: more jobs in the future than in the year of reference: it implies that the policy has a good impact on the situation under analysis;

- $\Delta_{jobs} < 0, y < Yr$: less jobs in the past than in the year of reference: it implies that the policy has a good impact on the situation under analysis;
- $\Delta_{jobs} < 0, y > Yr$: less jobs in the future than in the year of reference: it implies that the policy has a bad impact on the situation under analysis.

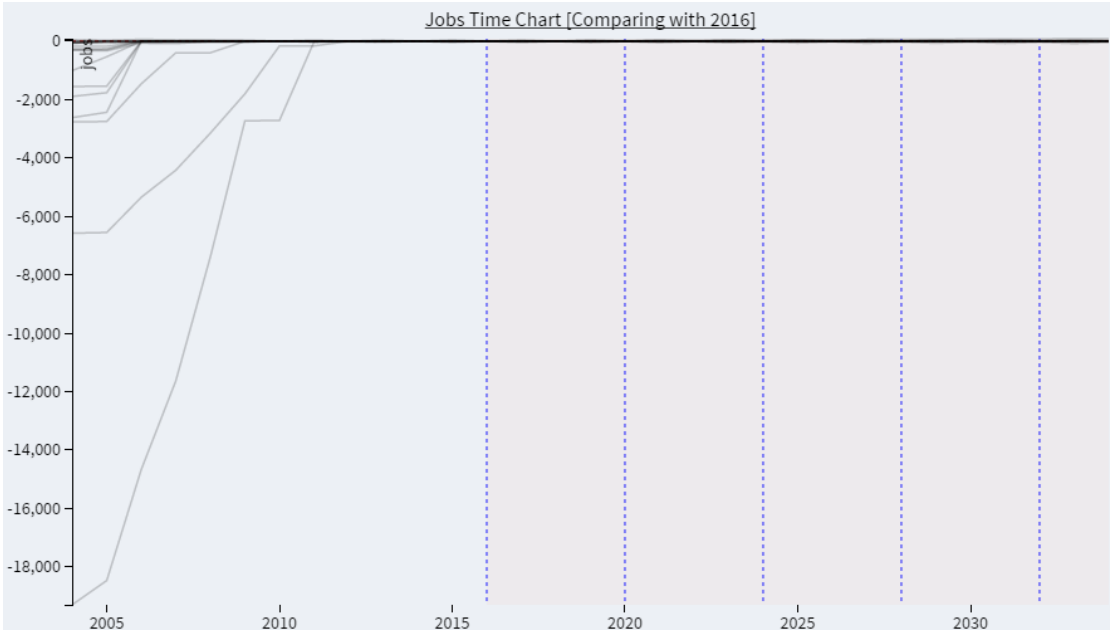

Figure 32 – Comparison of the values for the job indicator (Δ_{jobs} as in the equation at page 48) with respect to the reference year 2016: this chart highlights differences in the indicator values and allows to appreciate variations in the temporal trends. The data used to prepare this chart are the same as those in Figure 26 and refer to the districts of Madrid

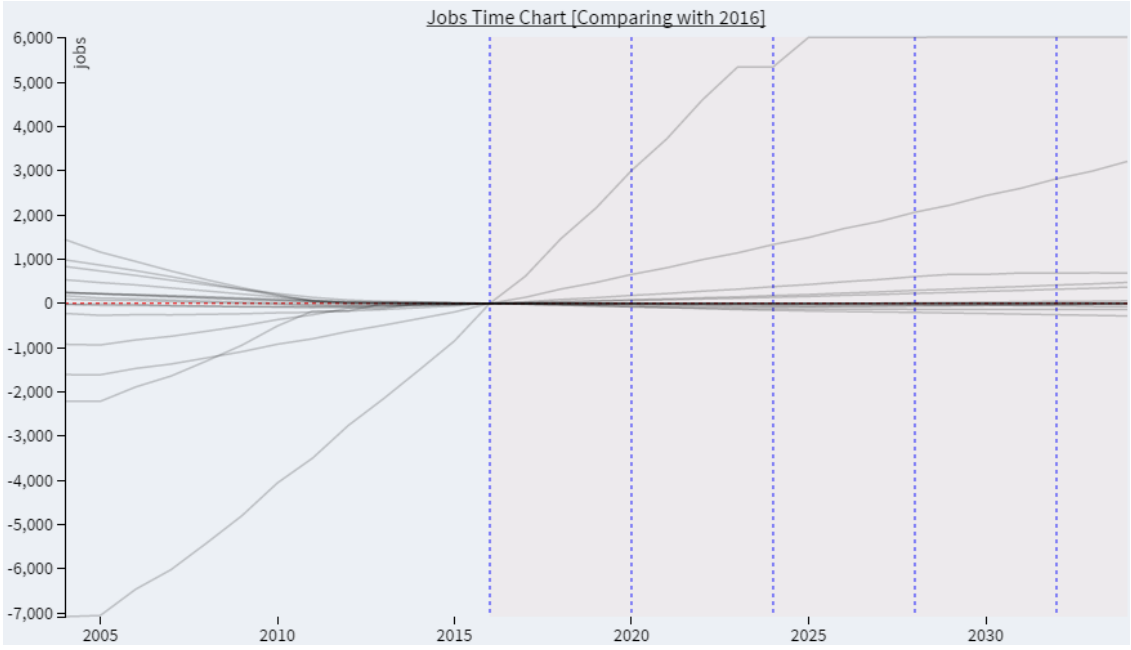

Figure 33 - Comparison of the values for the job indicator (Δ_{jobs} as in the equation at page 48) with respect to the reference year 2016: this chart highlights differences in the values of the indicator and allows to appreciate variations in the trends. The data used to prepare this chart are the same as those in Figure 27 and refer to the Northern part of the Metropolitan Region of Madrid

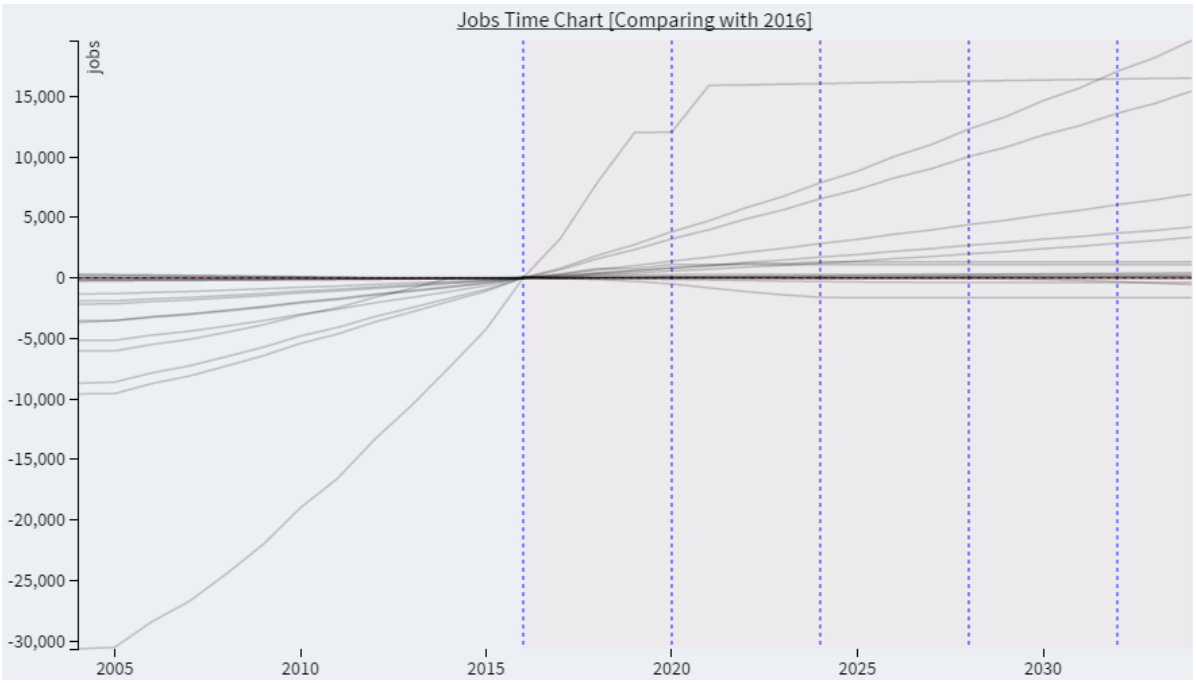

Figure 34 - Comparison of the values for the job indicator (Δ_{jobs} as in the equation at page 48) with respect to the reference year 2016: this chart highlights differences in the indicator values and allows to appreciate better variations in the trends. The data used to prepare this chart are the same as those in Figure 29 and refer to the Southern region part of the Metropolitan Region of Madrid.

The interpretation to give to such cases is much clear and straightforward by looking at the charts depicted in Figures from 32 to 34. In the remaining of this section we focus on the results obtained for the districts of Madrid and for the regions in the North and South, since these are the most suitable to illustrate the identification of relevant patterns. Figure 32 shows how the city of Madrid, according to the M40 scenario, reaches a sort of equilibrium with respect to the job values for the different zones towards the years 2011-2012 (there are actually small fluctuations of around hundreds of jobs for some zones starting from the years 2011-2012; further filtering and/or tooltips could help to discover this pattern).

It is interesting to note how especially the districts of Fuencarral – El Pardo and Barajas have performed a great jump in recovering their unfavourable situations in few years, and gaining, respectively, about 20000 and 6000 jobs. As previously noted, in the North of the Metropolitan Region of Madrid there are some zones showing negative trends at the beginning of the period of reference for the simulation. However, in Figure 27 this fact is hardly visible. On the contrary, Figure 33 presents a clear picture of this evidence: in particular, it is possible to appreciate how many regions are affected by this phenomenon and which is its magnitude. At the same time, it is remarkable the situation occurring in Tres Cantos, where the total number of jobs is expected to grow in 13000 along a 20-year period. Also, the situation in the southern part of the region of Madrid was chosen (and displayed in Figure 34) because, even if it tells a story very similar to the previous one, it highlights a negative trend that was not detected in the previous discussion. The zone affected by this problem is the one surrounding the town of El Soto, at the border with the autonomous region of Castilla La Mancha, and its magnitude is between 1300 and 1600 jobs. It could be considered a small quantity in comparison to other magnitudes revealed in this section, but the important socio-economic implications associated to a local decrease in the number of jobs makes the effect worth to analyse.

However, we have limited our analysis to explore the data of a single alternative. What about a comparison about the effects of the M40 cordon toll policy and the baseline scenario? To answer to this question, it is possible to refer to the choropleth maps displayed in Figure 35 and Figure 36, both showing interesting patterns for the period 2015-2021. Indeed, the dashboard is not only able to show the raw values of the simulator data, but moreover it can compare two scenarios in order to represent the benefits / drawbacks of different alternatives. Onto a map, it is also possible to detect the geographical patterns and easily spot which regions are forecasted to carry the most benefits / problems. With respect to the investigation carried out in this section, it is possible to evaluate whether, in a given period of reference, the toll would produce a decrease in the job indicator values compared against a no-intervention policy.

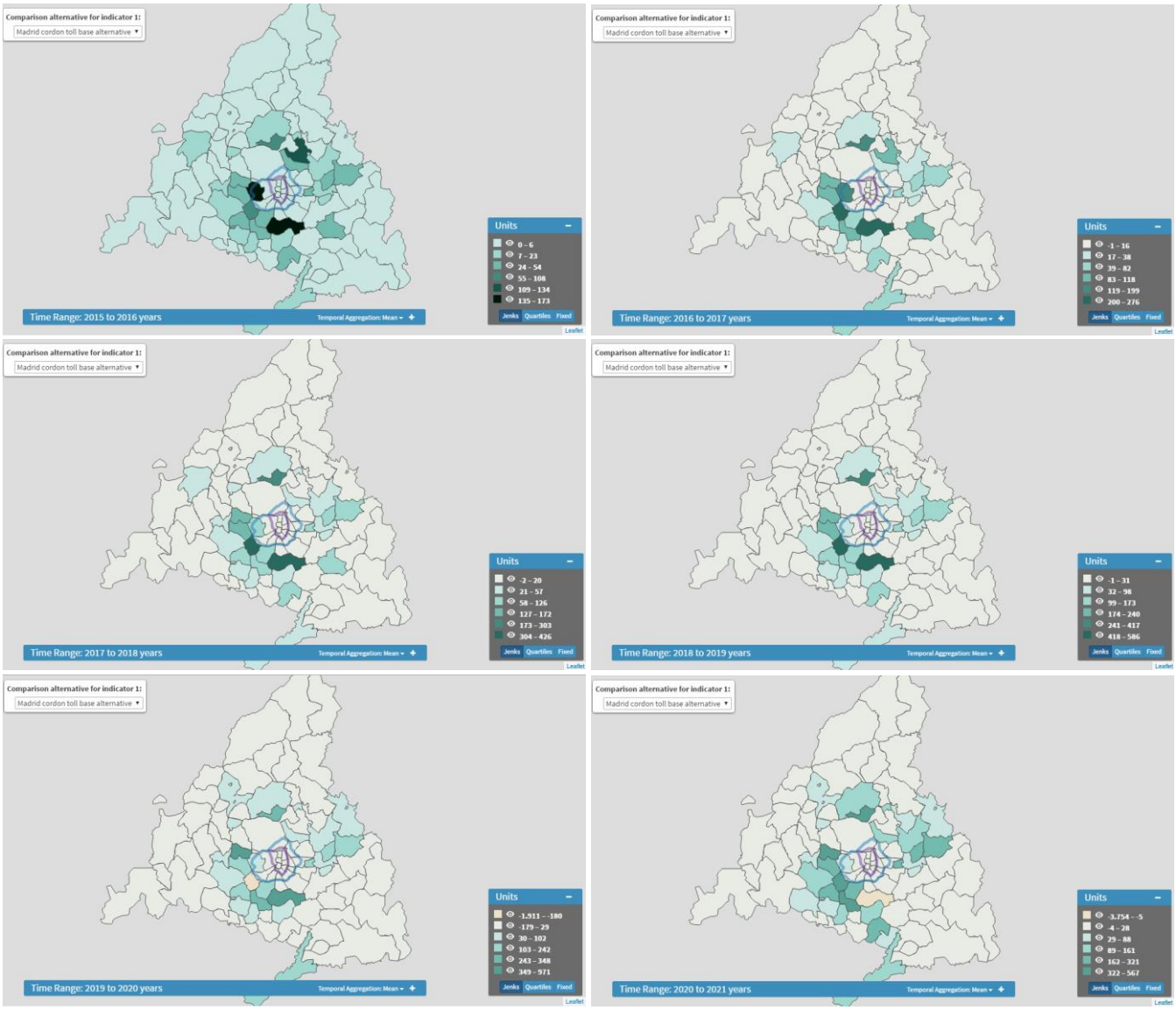

Figure 35 - Comparison of effects on the job indicator of the M40 cordon toll policy against the baseline scenario in the period 2015-2021. For the prevelance of green shades colouring the regions, the introduction of a cordon toll in the M40 beltway seems to have a positive effect on employment, especially in the towns immediately sorrounding the city of Madrid. The effects are more evident in the first year (leftmost image in the first row). A trend reversal can be spotted in the last two images, involving the towns of Alcorcón and Getafe.

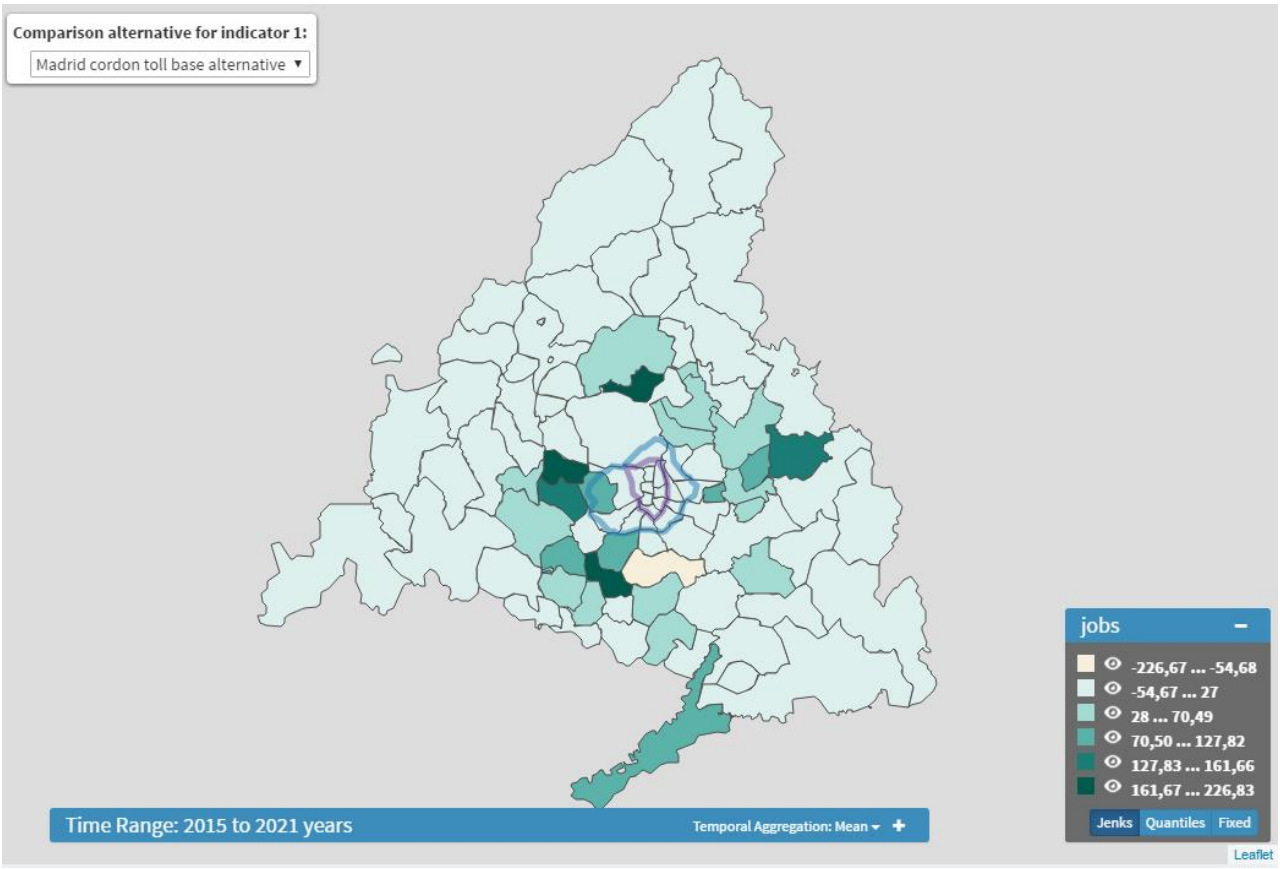

Figure 36 – Average variation in the number of jobs per zone across the region of study for the period of reference (2015-2021) resulting from the application of the M40 cordon toll policy versus the baseline scenario.

As a first insight, it can be seen that there is a positive effect in almost every zone during the six-years period analysed, since the regions are coloured in green. As a reminder, please note that: i) the choropleth map is now depicted with a diverging scale of colours since we want to make a comparison; in the default diverging scale, browns are used to represent negative values, greens for positive ones and greyish shades for those cases of equilibrium; ii) the comparison is made such that:

$$\Delta_{jobs} = jobs_{M40}^{z_i} - jobs_{baseline}^{z_i}$$

And, therefore, a positive value means that total number of jobs foreseen for the *i*-th zone for the M40 scenario is greater than the amount that is expected to result from the evolution of the situation without any policy intervention. The positive effects are more noticeable in the first year of the analysis and, once again, in most of the towns in the periphery of Madrid. Actually, these are the zones located just outside the M40 ring, which seems to support our hypothesis that the policy measure results into the creation or movement of job places to the periphery to avoid the toll. This could be considered an impulsive reaction, since the effects are mitigated little by little in the subsequent years. Nevertheless, the same regions are the most affected by the changes introduced by the M40 cordon toll policy: indeed, even if the raw differences are getting lower (clearer shades of green), the proportion between the best and worst zones is envisaged to remain approximately the same. Conversely, the districts of Madrid do not show any relevant change in the number of jobs since the related indicator is still positive and aligned to the general trend. When considering the results for the years 2019 and 2020 (last row of images in Figure 35), the attention of the analyst goes to a couple of regions, Alcorcón and

Getafe (both in the South of Madrid), where negative trends are depicted. However, these trends emerge for one year only and in both cases they turn to a positive value for the rest of the period considered in the simulation.

Finally, rather than considering specific differences year by year, let us have a look to the mean of the differences along the period of reference in order to have a general overview of the lustrum considered. The general picture derived from this analysis (and shown in Figure 36) is in line with the findings discussed in the previous paragraph. Therefore, it is worth taking a close look to the only relevant deviation, which is revealed by the fact that all the regions are coloured in green but one, corresponding to the city of Getafe. How is it possible if the same region is represented among the best performers almost every year (but in 2020)? To explain this fact, it is important to look at the legend, where each range is explained through its lowest and biggest value. This way, the analyst can see that the positive trends have an order of magnitude ranging between few dozens and some hundreds, while in the last image of Figure 35 the most negative value is about -3700. Then, the mean for Getafe is heavily affected by this last value, so that the final result turns to be negative. It would be interesting to perform further analysis to understand the causes of such a behaviour and to figure out possible corrections to overcome the problem. On the other side, this finding highlights one more time the importance of contrasting the evidence depicted by a single chart with different points of view to have a more general understanding of the big picture according different features. And having a tool allowing to perform such kind of analysis is one of the most important contribution of the platform developed in WP6.

4. Conclusions and future work

INSIGHT has created a visualisation platform aimed at supporting the assessment of policy options and the related decision making in urban planning processes. The platform relies on the visual analytics approach, provides interactive means for an active exploration and analysis of the data, and is built to provide insights about spatio-temporal and multi-dimensional features. As integral part of the three complementary deliverables produced in WP6, the emphasis of this document is on the most innovative contribution of INSIGHT in the field of visual analysis tools supporting policy making: the INSIGHT scatterplot. To illustrate the operational use of the different tools integrated into the platform, we present an example using the outputs of the cordon toll policy in Madrid.

Two main advances are discussed. Section 2 is more research-oriented since it focuses on the scientific contribution of a rather new visual technique for data representation, namely, polygonal scatterplots that are used to display multidimensional data on a 2D space. The multi-dimensionality of data is a critical aspect in planning processes because of the number of (heterogeneous) indicators to take into account when it is time to analyse a situation, characterise a geographical space to intervene on, and to define the objectives of a policy to monitor and improve. Therefore, it is crucial to have some chart to reason about them and look for patterns and relationships that are fleeing across multiple dimensions. The original idea of this chart is due to the work by Chen and colleagues (Chen et al., 2013). Starting from there, we implemented a version that could be able to extend some of the original functionalities. In particular, the main efforts were paid in investigating how different points placement strategy could help to explore data properties, discover hidden patterns and find trade-offs; and in adding interactive features to give to the analyst more context and reveal numerical properties of the data. Since the mapping from a *n*-dimensional to a bi-dimensional space inherently carries a loss of information as well as a flattening in the data distribution, some aspects about the interpretative problems that could arise by using such a chart have been presented as well as possible solutions to overcome these problems. In particular, it has been found that a major problem in understanding data patterns is linked with polygon having more than 4 sides. Nevertheless, the practical application of this visualisation to the simulation data has highlighted its potential and, at the same time, pointed out the directions to take to perform further research. In particular, it seems very interesting to perform a user study about topics such as perception, friend-usability, design choices and usefulness in accomplishing specific tasks, that could help to reveal unnoticed weaknesses or highlight new strong points.

On the other hand, Section **¡Error! No se encuentra el origen de la referencia.** has been conceived as a showcase on how it is possible to use the visualisation platform in its operational context. For this reason, we illustrated a practical example about how to get insights from data, taken from the case studies elaborated in the project. Of course, the ultimate goal of that section was not to draw conclusions about policies (as it could be done by a true analyst when interacting with the dashboard); it was rather to show how to coordinate the different data views to get a complete picture of the scenario; to highlight the benefits of a representation in a given context; and to take advantage of the different functionalities according to the analysis question in mind. While describing some remarkable evidence about Madrid data, the focus of the presentation was intended to remark the usefulness of the general approach followed during the design phase (see D6.1) and that could be summarised as follows:

- To provide a common, user-steering space to visually understand and compare policy formulations by showing different analytical points of view through a coordinated set of visualisations (e.g. multiple linked views, filters, …)

- To intercept the three main data features – geography, time evolution and multi-dimensionality – and present them at a glance;
- To provide proper interactive charts to reason about temporal trends, spatial patterns, characterisation of regions across multiple dimensions and trade-offs;
- To support operational tasks such as lookup, comparison, relation-seeking, and pattern search.

Finally, it is worth highlighting some more general conclusions going beyond the technical aspects of the implementation and more related to the general context of the project. First, the INSIGHT visualisation platform - even presenting improvable characteristics (such as more analytical functionalities or charts to display data) – has accomplished a number of objectives to ease the use of simulation data throughout the decision making process. In particular, the platform has been conceived and implemented as a connector of different simulators. As a unified space, it can be seen as both a showcase for policy alternative results as well as an analysis workspace; moreover, by dealing with the common features of each simulator data, it tries to set a bridge among heterogeneous tools: comparisons among simulator results are still and largely unfeasible, but at least it has been shown that their output can be analysed and visualised in a unified manner. This last evidence has been possible because of the efforts spent along the whole data chain, covering almost all the aspects such as the collection/importation to a common repository, the creation of interfaces for the (statistical) analysis and the use of common charts for the visualisation of the data of interest. According to this point of view, it is also safe to say that the platform could be seen as the translator connecting the data world with their final users (i.e. analysts, stakeholders, and policy-makers, …). Indeed, the use of suitable visual representations sets a common language of reference facilitating the exchange of information among the (human) actors involved in the process. And the use of interactive interfaces and visual metaphors is helpful to bring to the light data evidence in an active manner, then facilitating the general comprehension of the possible benefits and drawbacks of different policy options.

In the future, several enhancements could be foreseen. For instance, by providing tools being able to analyse and represent the Pareto frontiers of the whole policy space; by implementing specific analysis or graphical capabilities depending on the peculiarity of each simulator / policy to study; by expanding the number of simulators whose outputs are treated by the platform to reach a wider sample of city cases; and/or by equipping those already served with an interface with the tools required to (fully) automate the collection and pre-processing phases in the data process chain.

Annex I. Nomenclature notes

Throughout this document, and without loss of generality, we adopted the following norms to refer to some specific terms:

1. The term *user* refers as a general category of different professional figures, including but not limiting to, *analysts*, *policy makers* and *stakeholders*, as the main targets of the visual platform.
2. The words *dimension*, *objective* and *variable* are interchangeably used to refer to the same concept, namely one of the aspects or features a policy is founded on and evaluated. However, depending on the context, *dimension* has been mainly employed when talking about charts, *objective* about policy formulations and *variable* when treating their mathematical issues.

Annex II. References

Chen, S., Amid, D., Shir, O. M., Limonad, L., Boaz, D., Anaby-Tavor, A., & Schreck, T. (2013). Self-organizing maps for multi-objective Pareto Frontiers. *IEEE Pacific Visualization Symposium*, 153–160. doi:10.1109/PacificVis.2013.6596140

Gavin, D. G., Oswald, W. W., Wahl, E. R., & Williams, J. W. (2003). A statistical approach to evaluating distance metrics and analog assignments for pollen records. *Quaternary Research*, *60*(3), 356–367.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. doi:10.1109/5.58325

Korhonen, P., & Wallenius, J. (2008). *Visualization in the multiple objective decision-making framework*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5252 LNCS). doi:10.1007/978-3-540-88908-3-8

Lotov, A. V., & Miettinen, K. (2008). Visualizing the pareto frontier. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5252 LNCS*, 213–243. doi:10.1007/978-3-540-88908-3-9

Taneja, I. J. (1989). On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, *76*, 327–413.

Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2). Graphics press Cheshire, CT.