Pontifícia Universidade Católica
do Rio de Janeiro

**Guilherme Gonçalves Schardong**

# Visual interactive support for selecting scenarios from time-series ensembles

**Tese de Doutorado**

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática.

Advisor : Prof. Hélio Côrtes Vieira Lopes
Co-advisor: Profª. Simone Diniz Junqueira Barbosa

Rio de Janeiro
September 2018

Pontifícia Universidade Católica
do Rio de Janeiro

**Guilherme Gonçalves Schardong**

# Visual interactive support for selecting scenarios from time-series ensembles

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática. Approved by the undersigned Examination Committee.

**Prof. Hélio Côrtes Vieira Lopes**
Advisor
Departamento de Informática – PUC-Rio

**Profª. Simone Diniz Junqueira Barbosa**
Co-advisor
Departamento de Informática – PUC-Rio

**Prof. Fernando Luiz Cyrino Oliveira**
Departamento de Engenharia Industrial – PUC-Rio

**Prof. Bruno Fânzeres dos Santos**
Departamento de Engenharia Industrial – PUC-Rio

**Prof. Regis Kruel Romeu**
CENPES

**Prof. Sergio Lima Netto**
UFRJ

**Prof. Abelardo Borges Barreto Junior**
Departamento de Matemática – PUC-Rio

**Prof. Alex Laier Bordignon**
UFF

**Prof. Márcio da Silveira Carvalho**
Vice Dean of Graduate Studies
Centro Técnico Científico – PUC-Rio

Rio de Janeiro, September $13^{th}$, 2018

**Guilherme Gonçalves Schardong**

Bachelor's in Computer Science (2011) and Masters' in Informatics (2014) at the Federal University of Santa Maria (UFSM), with emphasis on Applied Computing. Worked for the Applied Computing Laboratory (LaCA) from 2009 to 2014 and Tecgraf from 2014 to 2016 in the Reservoir visualization group. Since 2016 works at GALGOS, former ATD-Lab, at PUC-Rio.



Bibliographic data

Schardong, Guilherme Gonçalves

Visual interactive support for selecting scenarios from time-series ensembles / Guilherme Gonçalves Schardong; advisor: Hélio Côrtes Vieira Lopes; co-advisor: Simone Diniz Junqueira Barbosa. – Rio de janeiro: PUC-Rio, Departamento de Informática, 2018.

v., 92 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia

1. Redução de Cenários. 2. Interatividade. 3. Visualização Científica. 4. Tomada de Decisão. 5. Conjuntos de Dados Temporais. I. Lopes, Hélio Côrtes Vieira. II. Barbosa, Simone Diniz Junqueira. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

## Acknowledgments

To God, for giving me the strength to endure all sacrifices during these last four years, for the blessings, friends and brothers found during this journey.
To my parents, José Luiz and Rosangela for the care, support, friendship and unconditional love. You have my eternal thanks.
To Joseany, my girlfriend, for showing me places I will never forget. For trailing this path with me, sharing all the joys and enduring the hardships along the way. I will never be able to fully repay you, all that I can give you is my eternal love and friendship.
To profs. Hélio Lopes and Simone Barbosa for the help and advice. If I arrived here today, it was because of your counsel and help along the way.
To prof. Waldemar Celes for the job opportunity at Tecgraf/PUC-Rio and mentoring during my first year.
To the visualization group at Tecgraf for all lessons, laughs and growth opportunities.
To Fabrício Cardoso for the "Halo nights" on Saturdays, the counsels, talks and for showing me all that I know about Pará's culture and customs. What a wonderful place.
To GALGOS and all the crazy lunatics that I met while working there. This journey would certainly be boring without you people.
To all others not mentioned here. Don't think I have forgotten you, you were all crucial to me during this time, and will continue to be so in future journeys.
To Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for partially financing this research under grant 153737/2014-0. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

# Abstract

Schardong, Guilherme Gonçalves; Lopes, Hélio Côrtes Vieira (Advisor); Barbosa, Simone Diniz Junqueira (Co-Advisor). **Visual interactive support for selecting scenarios from time-series ensembles**. Rio de Janeiro, 2018. 92p. Tese de doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Stochastic programming and scenario reduction approaches have become invaluable in the analysis and behavior prediction of dynamic systems. However, such techniques often fail to take advantage of the user's own expertise about the problem domain. This work provides visual interactive support to assist users in solving the scenario reduction problem with time-series data. We employ a series of time-based visualization techniques linked together to perform the task. By adapting a multidimensional projection algorithm to handle temporal data, we can graphically present the evolution of the ensemble. We also propose to use cumulative bump charts to visually compare the ranks of distances between the ensemble time series and a baseline series. To evaluate our approach, we developed a prototype application and conducted observation studies with volunteer users of varying backgrounds and levels of expertise. Our results indicate that a graphical approach to scenario reduction may result in a good subset of scenarios and provides a valuable tool for data exploration in this context. The users liked the interaction mechanisms provided and judged the task to be easy to perform with the tools we have developed. We tested the proposed approach against state-of-the-art techniques proposed in the literature and used in the industry and obtained good results, thus indicating that our approach is viable in a real-world scenario.



## Keywords

Scenario Reduction; User Interaction; Scientific Visualization; Decision Making; Time Series Ensembles.

## Resumo

Schardong, Guilherme Gonçalves; Lopes, Hélio Côrtes Vieira; Barbosa, Simone Diniz Junqueira. **Uma Abordagem Visual e Interativa para a Seleção de Conjuntos de Cenários Temporais**. Rio de Janeiro, 2018. 92p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O uso de abordagens de programação estocástica e redução de cenários tem se tornado imprescindível na análise e predição de comportamento de sistemas dinâmicos. Entretanto, tais técnicas não levam em conta o conhecimento prévio sobre domínio que o usuário possui. O presente trabalho tem por objetivo o desenvolvimento de uma abordagem visual e interativa para abordar o problema de redução de cenários com dados temporais. Para tanto, nós propomos a implementação de uma série de visualizações de dados temporais integradas. Também propomos a adaptação de um algoritmo de projeção multidimensional para lidar com dados temporais. Desta forma, podemos representar graficamente a evolução de um conjunto de cenários ao longo do tempo. Outra visualização proposta no presente trabalho é uma adaptação de *Bump chart* para lidar com dados temporais acumulados; através dele, um usuário pode comparar a evolução das distâncias entre os diferentes cenários e um cenário de referência. Para validar a nossa proposta, fizemos uma implementação das técnicas propostas e conduzimos um estudo com usuários de diferentes áreas do conhecimento e níveis de experiência. Os resultados obtidos até então indicam que uma abordagem visual para o problema de redução de cenários é viável, e permite a seleção de um conjunto razoável de cenários. Além disso, constatamos que essa abordagem pode ser útil em um contexto de exploração de dados visando a redução de cenários. O usuário também pode explorar visualmente os resultados de outras técnicas de redução de cenários usando nossa abordagem. Os usuários entrevistados reportaram facilidade em cumprir as tarefas propostas e comentaram positivamente sobre os mecanismos de interação fornecidos pelo nosso protótipo. Também testamos os cenários escolhidos usando nossa proposta contra outras abordagens encontradas tanto na literatura quanto em uso na indústria. Os resultados obtidos foram bons, indicando que nossa proposta é viável em casos de uso reais.

## Palavras-chave

Redução de Cenários; Interatividade; Visualização Científica; Tomada de Decisão; Conjuntos de Dados Temporais.

# Table of contents

## List of figures

# List of tables

## List of abreviations

ANP – Agência Nacional do Petróleo
CENPES – Centro de Pesquisas Leopoldo Américo Miguez de Mello
LAMP – Local Affine Multidimensional Projection
MDS – Multidimensional Scaling
MP – Multidimensional Projections
$N_p$ – Cumulative oil production
$P_{10}$ – Percentile 10
$P_{50}$ – Percentile 50
$P_{90}$ – Percentile 90
PETROBRAS – Petróleo Brasileiro S.A.
$Q_o$ – Oil flow
$Q_w$ – Water flow
$W_p$ – Cumulative water production

*ALL THINGS THAT ARE, ARE OURS. BUT WE MUST CARE. FOR IF WE DO NOT CARE, WE DO NOT EXIST. IF WE DO NOT EXIST, THEN THERE IS NOTHING BUT BLIND OBLIVION. AND EVEN OBLIVION MUST END SOMEDAY. LORD, WILL YOU GRANT ME JUST A LITTLE TIME? FOR THE PROPER BALANCE OF THINGS. TO RETURN WHAT WAS GIVEN. FOR THE SAKE OF PRISONERS AND THE FLIGHT OF BIRDS.*
*LORD, WHAT CAN THE HARVEST HOPE FOR, IF NOT FOR THE CARE OF THE REAPER MAN?*

**Terry Pratchett**, *Reaper Man.*

# 1
# Introduction

Recent developments in simulation techniques have helped researchers to better understand and predict several naturally occurring phenomena, ranging from weather forecast (Sanyal2010) to circuit calibration (Lee2010) and fluid dynamics (Hummel2013). These simulations produce a huge amount of data, due to the availability of computing power and simulation model refinements. To extract meaningful information from all those data, researchers have been developing an array of approaches in diverse areas: data mining (Wang2014), machine learning (Yang2007), visualization (Phadke2012), and optimization (Alrefaei2007). Such approaches typically use statistical measures to summarize the results or to reduce the dimensionality of data and select the most probable outcomes of the simulation.

In simulation analysis, scenario reduction is particularly useful, since its goal is to reduce the number of simulation outcomes (i.e., *scenarios*) to a more manageable size, with minimal loss of variability. Existing approaches are usually modeled as stochastic programming problems (Dupacova2003, Armstrong2013), in which a probability is associated to each possible scenario, and the goal is to select a subset of scenarios whose probability is closest to that of the original set (henceforth called *ensemble*).

However, none of those approaches actively engages the users and their knowledge about the problem domain. Visualization-based approaches, conversely, allows for interactive exploration of the data through visual tools and interaction mechanisms. Visual analytics supports decision making by integrating the best of computational processing power and human cognitive prowess (Aigner2007, Andrienko2011, Keim2008, Kohlhammer2011). For time-based ensembles, (Cheng2016) provide a comprehensive survey on time-series and time-based visualization techniques and interaction mechanisms, from which we draw in our proposal.

The main goal of our work is to provide visual interactive support for solving the scenario reduction problem with time-series data. We employ a series of time-based visualization techniques linked together, allowing the user to draw from the strengths of each technique. To the best of our knowledge, no one has proposed a similar way of approaching this problem.

We also propose adaptations to two known visualization algorithms: (i) the Local Affine Multidimensional Projection (LAMP) algorithm (Joia2011), in order to produce a time-based representation of the data; and (ii) Bump charts, also known as Slope graphs (Tufte1990), in order to view a transformed version of our time-series ensemble. Multidimensional projections are evolving, especially in the works of (Alencar2012) and (Wong2013). By using a different base technique, with a strong mathematical foundation (Joia2011), we aim to provide a more robust representation of the similarity between the series over time. We have also made some adaptations to Bump charts (Tufte1990). As our data is not ordinal, we transform them by ranking the distance between each series and a baseline. Moreover, we do not treat each time step as isolated from the others, but as an accumulated rank from all previous time steps.

As proof of concept, we built a prototype software using the brushing and linking framework, proposed by (Becker1987, Buja1991), as basis for the user interaction with the different visualizations. We chose four visualization mechanisms: (i) a Fanchart, proposed by (Britton1998); (ii) the Distance scatterplot; (iii) a cumulative Bump chart; and (iv) a scatterplot with the results of our proposed multidimensional projection (MP) approach. To evaluate our approach, we conducted an empirical study involving experts and non-experts in the scenario reduction area.

In summary, our contributions are:

- A visual interactive approach to assist the user in selecting a subset of meaningful scenarios from a time-series ensemble dataset, thus solving an instance of the scenario reduction problem;
- An adaption of a multidimensional projection algorithm to generate a visual representation of time-varying data, taking into account the time component of the data;
- A transformation of a time-series ensemble dataset into a cumulative, ranked version, in order to support a visual assessment of its evolution.

The main motivation for our work came from a project proposed by PETROBRAS/CENPES in partnership with PUC-Rio. The project is entitled: "*Visualização e quantificação de incertezas de um conjunto de simulações de reservatório*", with a registry number 18008-3 on the National Petroleum Agency (ANP in portuguese). During the course of this project, researchers from CENPES raised the percentile selection problem as one of the main issues faced during the decision making process for oil reservoir management. During our research, we found that this problem is an instance of the scenario reduction

problem mentioned above, thus techniques used in this area could, and were applied to the percentile selection problem (Shirangi2016, Meira2016, Sarma2013, Scheidt2009a, Scheidt2009b, Armstrong2013, Armstrong2014). However, none of these approaches took the evolution of the reservoir into consideration, leaving a gap in the literature which we aimed to fill.

## 1.1 Thesis Structure

The remainder of this thesis is organized as follows: Chapter 2 presents related works grouped by topic: Scenario Reduction and Multidimensional Projection. Next, in Chapter 3 we describe our approach and explain the visualization techniques employed, and interaction mechanisms provided. Chapter 4 presents details about the experiments and the associated results. Finally, Chapter 5 presents some concluding remarks and directions for future work.

# 2
# Related Work

This chapter describes two groups of related works, on: (i) scenario reduction (i.e., the problem we address); and (ii) multidimensional projections, which are the type of algorithm we have adapted to help us with scenario reduction. Table 2.1 presents some of the main works and their contributions compared to ours. Details about these works are presented below.

Table 2.1: Overview of the literature in scenario reduction and multidimensional projections

| **Work** | **Scenario Reduction** | **Multidimensional Projections** | **Brushing & Linking** |
|---|---|---|---|
| (Armstrong2014) | X | | |
| (Growe-Kuska2003) | X | | |
| (Meira2016) | X | | |
| (Lee2010) | X | | |
| (Heitsch2003) | X | | |
| (DiDomenica2007) | X | | |
| (Kawas2014) | X | | |
| (Park2016) | | | X |
| (Demir2014) | | | X |
| (Scheidt2009b) | X | X | |
| (Sahaf2016) | X | | X |
| (Waser2014) | X | | X |
| Our approach | X | X | X |



## 2.1
## Scenario Reduction

As the number of objects in an ensemble grows, it becomes increasingly difficult to analyze or visualize it adequately, even when using dimensionality reduction techniques. In many cases, it becomes necessary to select a representative subset of the ensemble for further processing, in a process known as *scenario reduction*, which has increasingly attracted researchers' interest, especially in areas such as power production (Dupacova2003, Growe-Kuska2003) and geostatistics (Scheidt2009a, Scheidt2009b, Heitsch2009, Lee2010,

Sarma2013, Armstrong2013, Armstrong2014). A number of researchers have proposed to use stochastic programming as an approach to tackle this problem. (Dupacova2003) stated that the Fortet-Mourier family of probability metrics may be used as canonical metrics to find a subset of scenarios with probability distributions closest to the original set. They reduced the number of possible scenarios by 50% while keeping 90% relative accuracy in the remaining scenarios. More recently, (Armstrong2013) proposed a metric for the distance between conditional simulated realizations of ore deposits, along with a random search procedure to find an approximation of the ideal subset of scenarios. They followed the approach proposed by (Heitsch2009) to calculate the distance between a subset of scenarios and the full ensemble. In their experiments, the best subset found was 1% off the expected value for their objective function, which indicates that the number of possible scenarios can be strongly reduced without significant loss in variability.

In the petroleum field, (Scheidt2009b) have used dimensionality reduction and kernel methods to quantify the uncertainty in an ensemble of geological facies realizations. Their approach involved mapping the realizations onto a lower dimensional space using a multidimensional scaling (MDS) algorithm (Kruskal1978) and flow-related distance metrics, such as the Hausdorff distance (Suzuki2006) or time-of-flight-based metrics (Park2007). They have also used kernel methods to transform the projected points from a non-linear space onto a linear one, thus facilitating the application of grouping approaches, such as clustering algorithms and Principal Component Analysis. After defining the realization groups, a few elements of each group are chosen for the actual flow simulation. The flow simulation statistics they obtained with a reduced number of realizations were very similar to those with the full ensemble.

Different from most of the scenario reduction approaches presented so far, our main goal is to allow users to input their own knowledge of the problem domain into the process through graphical tools, therefore leading to a more flexible process overall. To the best of our knowledge, there is little work on visual analytics and scenario reduction. (Sahaf2016) proposed a scenario reduction approach based on randomly sampling scenarios after clustering them using a mutual information similarity metric. They implemented this approach in a visual analytics framework, where it is possible to visualize the spatial contribution of each model to the similarity of scenarios and run a clustering algorithm in an area specified by the user. (Kawas2014) proposed an uncertainty-aware framework for decision optimization, in which they employed classic stochastic programming to perform the scenario reduction and used visual analytics only to evaluate the resulting models.

On a decision support systems context we found no works that involve both scenario reduction and visual analytics. (Waser2014) comes close, by proposing a scenario generation and interactive visualization approach applied to flooding management. Their approach can simulate flooding scenarios in real time, but their visualization tool is not scalable and the plans generated are suboptimal. (DiDomenica2007) incorporate stochastic programming and scenario generation techniques into established decision support and information systems. They successfully argue that decision and simulation models can be combined in order to create business analytics, therefore creating uncertainty-aware decision and information systems. (Park2016) proposed a visual analytics approach for managing supply chain networks. They modeled these networks as directed graphs and implemented a series of interactive visualizations for them, including: force-directed layout, treemap layout, substrate-based visualization, chord diagrams and matrix layout. However, their views are not connected to each other, therefore lacking an important pattern-discovery mechanism.



## 2.2 Multidimensional Projections

Multidimensional Projection (MP) techniques help us explore complex datasets. Using adequate distance metrics and dimensionality, patterns in the data may stand out, allowing users to quickly identify them. Their usefulness motivated the development of new MP techniques and the adaptation of existing techniques to specific kinds of data and application domains.

(Wong2013) used MP techniques to explore time-varying volume data. They adapted two algorithms: Fastmap (Faloutsos1995) and Part-Linear Projection (Paulovich2010PLP), in order to preserve temporal coherence between data volumes at different time steps. The new algorithms, named Time-Coherent Fastmap (TC-Fastmap) and Time-Coherent Part-Linear Projection (TC-PLP), achieved an equivalent or lower configuration stress when compared to other time-varying projection techniques. They also proposed a scatter projection for attribute-space data exploration and for correlating selections to the object-space model.

(Alencar2012) adapted the Least Squares Projection (LSP) algorithm proposed by (Paulovich2008) to show the temporal evolution of groups of data. They applied their Time-based LSP algorithm to visualize the evolution of articles written by a researcher between the years 1995 and 2010. They plotted the results as a graph, with edges representing references between two articles, vertex color showing the publication age, and vertex size indicating the count

of citations to each paper by 2010. They also employed a DBSCAN clustering algorithm (Ester1996) to identify groups of similar papers and extracted the topics of each group using the approach presented by (Eler2009). In scientific paper collections, this is especially useful to assess the evolution of research topics of a knowledge area, and may help to identify and predict research trends.

The first part of our work draws on (Alencar2012): we propose an adaptation of the Local Affine Multidimensional Projection (LAMP) algorithm (Joia2011) to generate a sequence of mappings of our time series ensemble. Each time step results in a new projection of the data until then. By merging the results of those projections in a single view, we provide a graphical approach to assess the evolution and behavior of the ensemble.

# 3
# Background

We discuss in this chapter some fundamental concepts required to understand our work. We start by defining the concept of a baseline series in Section 3.1, then, we define the Fanchart, Distance and Bump charts in sections 3.2, 3.3 and 3.4 respectively. In Section 3.5 we provide a definition for the Local Affine Multidimensional Projection (LAMP) algorithm, that serves as a basis for the Time-lapsed Local Affine Multidimensional Projection adaption we propose, further detailed in Section 3.6. Finally, in Section 3.7 presents the interaction mechanisms implemented as well as further details about our prototype.



## 3.1
## Definition of a baseline percentile series

Given a set numbers, a percentile is a value that divides those numbers in two subsets, those with numerical value larger than it, and those with numerical value smaller than or equal to it. For example, in the set of numbers $X = \{1, 2, 2, 2, 3, 4, 5, 8, 8, 8, 9\}$, the percentile 0, or $P_0$ is $\min(X) = 1$, while the percentile 100 is $P_{100} = \max(X) = 9$. The percentile 50, or $P_{50}$ is a value larger than, or equal to half of values of $X$, in the example, $P_{50} = 4$.

Calculating the percentile values for single-variate distributions is a well-defined problem. While there are several different ways for calculating these values, this is a well-defined and well-resolved problem. Such statement is not true for multivariate distributions, where the percentile is not well-defined. In this work, we use time-series as baselines for selecting a subset of scenarios, we call these baselines "percentile series". However, they do not fit the definition of a percentile, since we cannot guarantee that each of them will be larger than, or equal to a certain percentage of series in the ensemble. Figure 3.1 shows a set of three time-series, where the percentile series is marked in blue circles. This series, which is a member of the ensemble, is a $P_{50}$ series, since it divides the ensemble in two subsets, those larger than it (Series 3), and those smaller, or equal to it (Series 1 and 2).

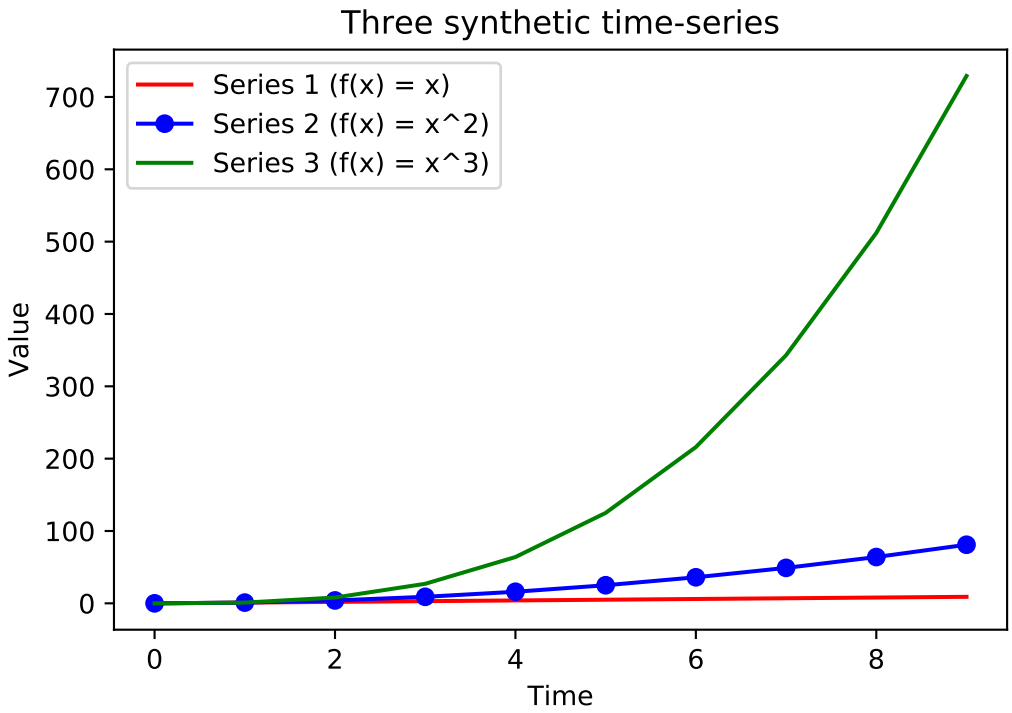
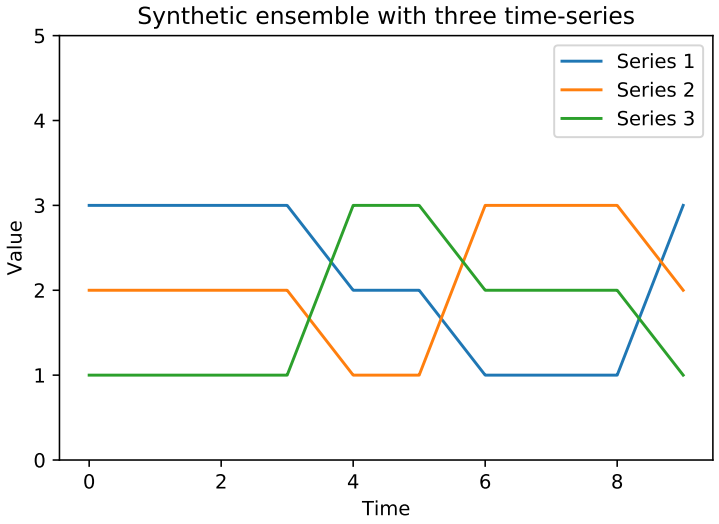
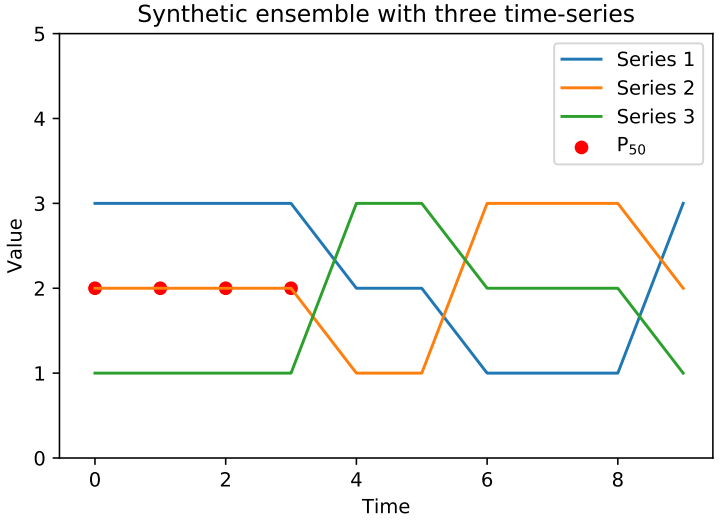



Figure 3.1: Synthetic time-series ensemble with three series, numbered 1 to 3 and the $P_{50}$ series marked in blue circles.

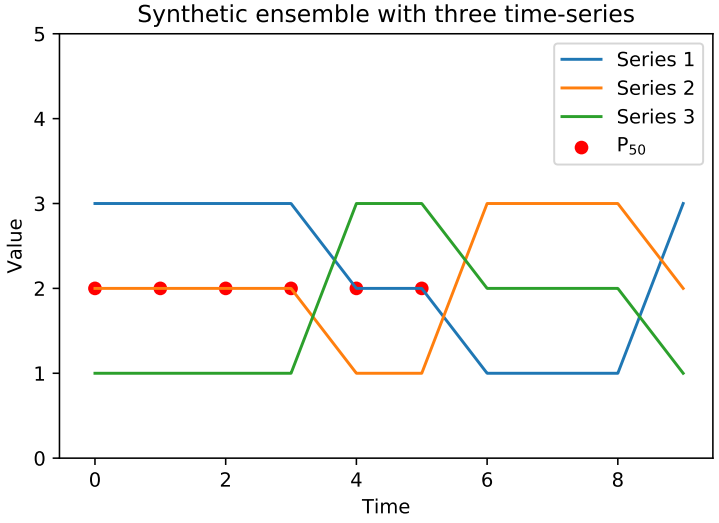
We build these percentile series by calculating the percentile values at each individual time, the series is then defined by connecting these individual percentile values. Figure 3.2 shows the process of building a $P_{50}$ baseline series for a synthetic time-series ensemble with three elements. Notice that the resulting $P_{50}$ series is a mix of the series in the original ensemble. At each individual time, the $P_{50}$ series is larger than, or equal to at least half the values, however, the resulting $P_{50}$ series is not a percentile according to the definition, since it does not fit the definition. For the purposes of this work, the connected percentile points will be called a percentile series and these series will be used as baselines for the selection process to be explained in this chapter.



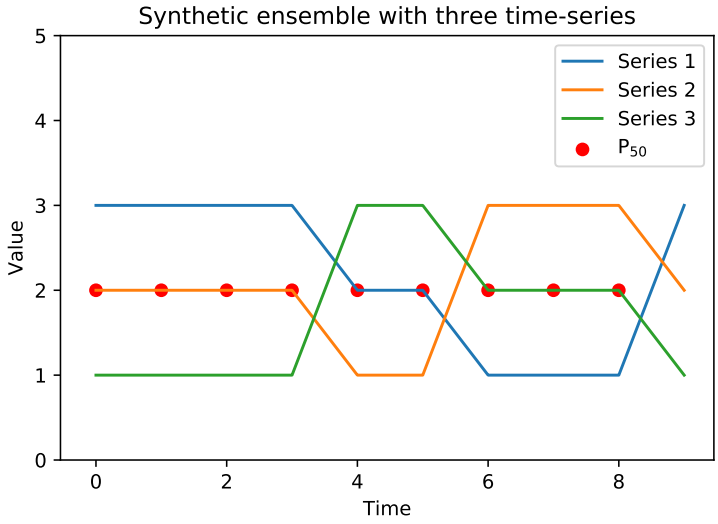

(a) Original ensemble with three synthetic time-series.



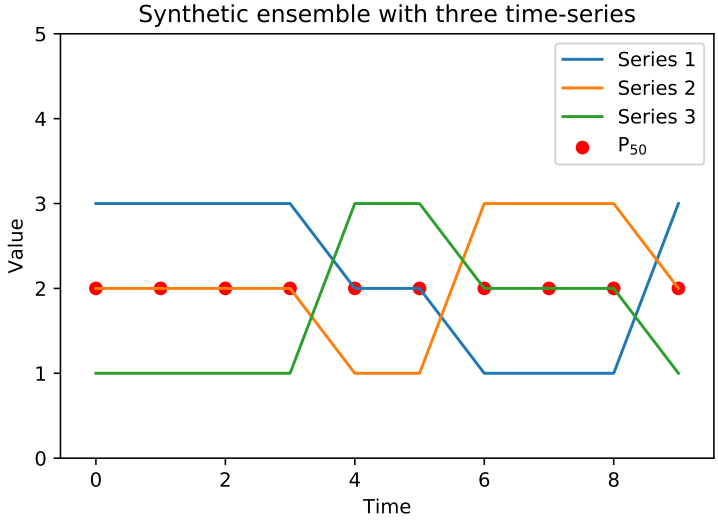

(b) $P_{50}$ values calculated for times $[0, 3]$. For these times, they coincide with Series 2.



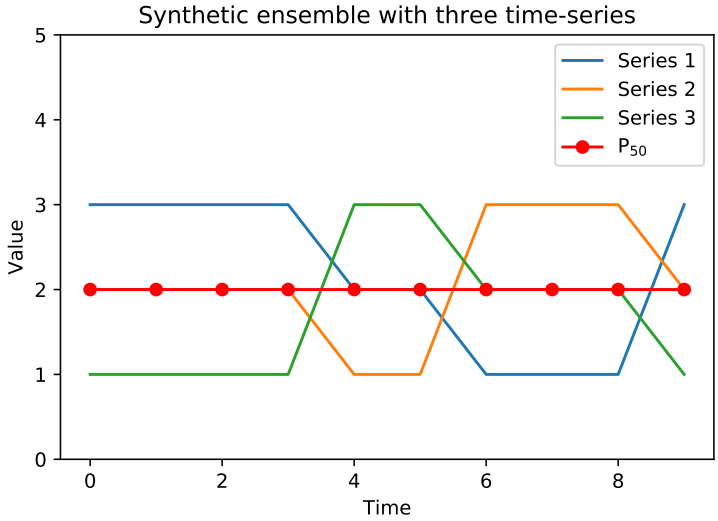

(c) $P_{50}$ values calculated for times $[0, 5]$. For times 4 and 5, they coincide with Series 1.




(d) $P_{50}$ values calculated for times $[0, 8]$. For times $[6, 8]$, the $P_{50}$ values coincide with Series 3.




(e) $P_{50}$ values calculated for the whole time range. In the last time, the $P_{50}$ value coincides with Series 2.



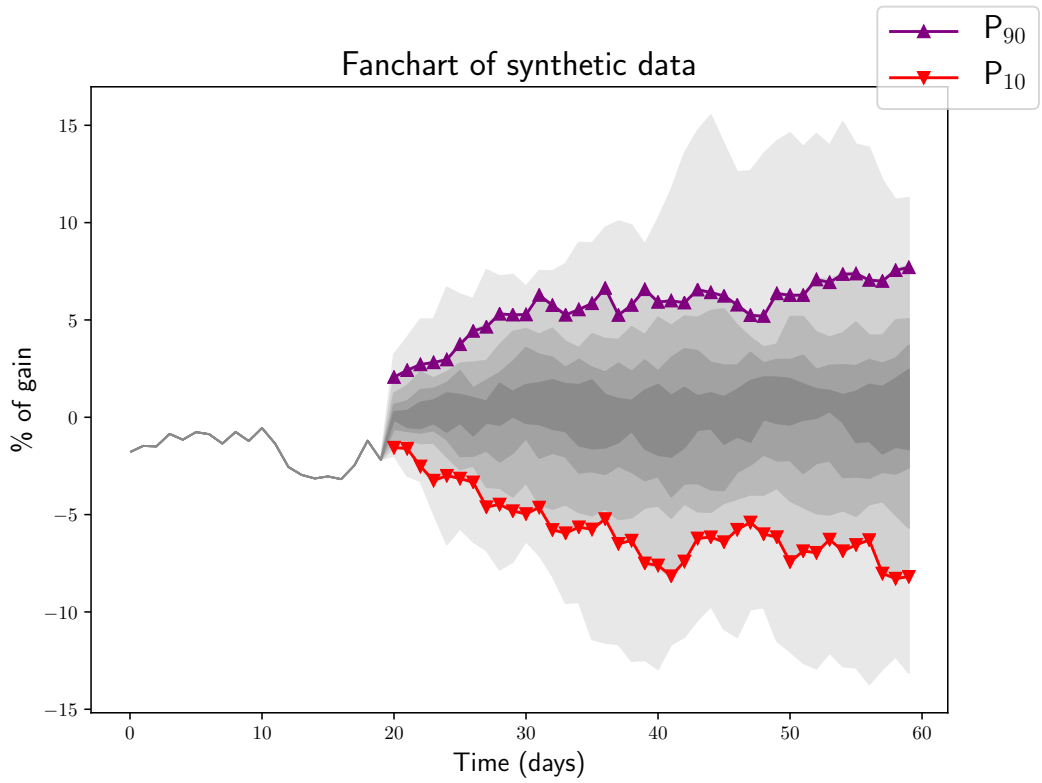

(f) $P_{50}$ series, built by connecting all $P_{50}$ values.

Figure 3.2: Building the $P_{50}$ percentile series.

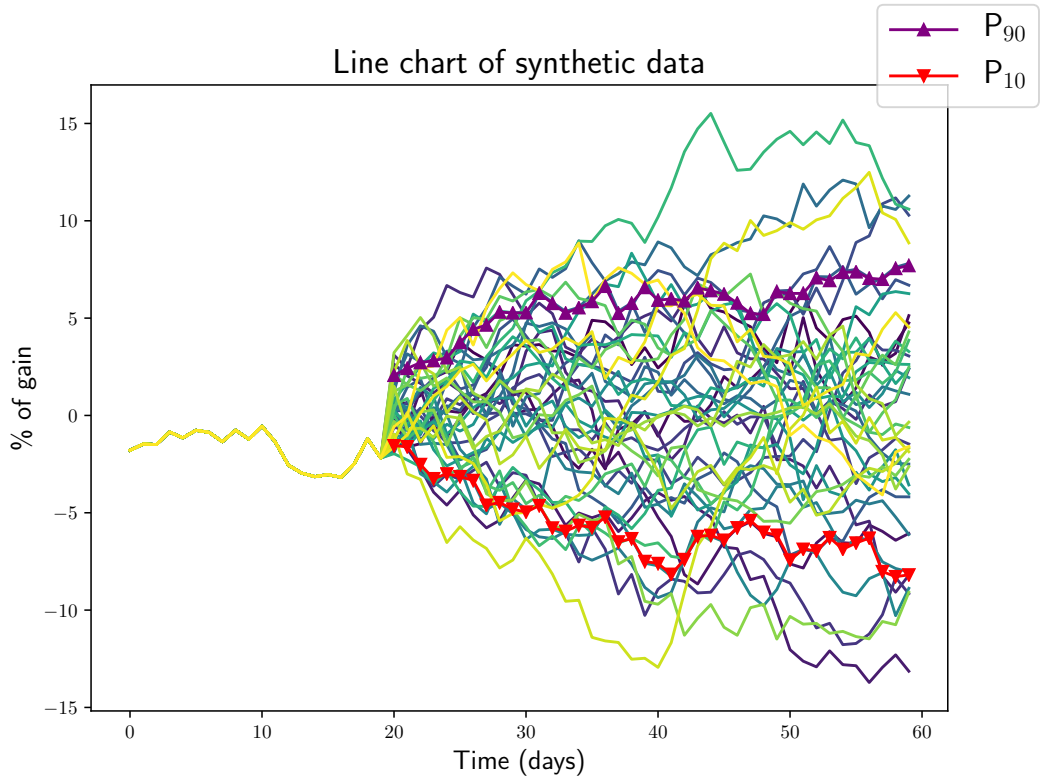
## 3.2 Fanchart

The first graphical view used in our approach is the Fanchart (Britton1998), used to visualize a distribution of time-based data. The Fanchart is commonly used to graph observed past data together with forecasts of future data. The observed data is represented as a line chart, since their values are known, while the forecasts are represented as an increasingly wide cloud of possibilities. Since the values around the mean are, usually, more likely to happen, they are represented in stronger colors. As the values stray further to the extremities, their color gets fainter, a reflection of the smaller likelihood of their occurrence.

Fancharts are useful in uncertainty analysis, since a wide fan of forecasts represents more uncertainty about the future, while a narrower fan represents less uncertainty. Compared to line charts, a Fanchart is less cluttered visually, and thus a good choice for assessing a large ensemble of time-series. In our approach, Fancharts are the closest view a user has of the raw data. When analyzing scenarios, analysts can use the Fanchart to check whether the behavior shown in the other views is consistent. Figure 3.3a shows an example of a Fanchart of a synthetic time-series ensemble, compared to a line chart of the same ensemble (3.3b).




(a) Fanchart




(b) Line chart

Figure 3.3: Visual comparison between a fanchart (3.3a) and line chart (3.3b) of a synthetic time-series ensemble. The $P_{10}$ and $P_{90}$ series are calculated and shown only for the simulated times.

The time-series ensemble used in the examples of this chapter contains forty elements, in addition to the $P_{10}$ and $P_{90}$ series. Each series contains twenty days of historic data and forty days of simulated data. Both the historic and simulated data are artificial, created using a random walks. A random walk $x(t)$, where $t$ is the current time, is defined as: $x(t) = x(t) + w(t)$, where $w(t)$ is white noise series at time $t$. The white noise series $w(t)$ was generated using a normal distribution with $\mu = 0, \theta = 1$. For the purposes of our examples in this chapter, we will treat the historic data as observed values from a naturally occurring process, and the simulated data as simulations performed to predict the behavior of this process.

## 3.3 Distance chart

The second graphical view is the Distance chart. As its name implies, the Distance chart graphs the distances, or similarities, between a set of objects and a baseline. Selecting the scenarios closest to a reference may provide a reasonable starting set of solutions to scenario reduction. Figure 3.4 shows an example of a Distance chart of a synthetic time-series ensemble.

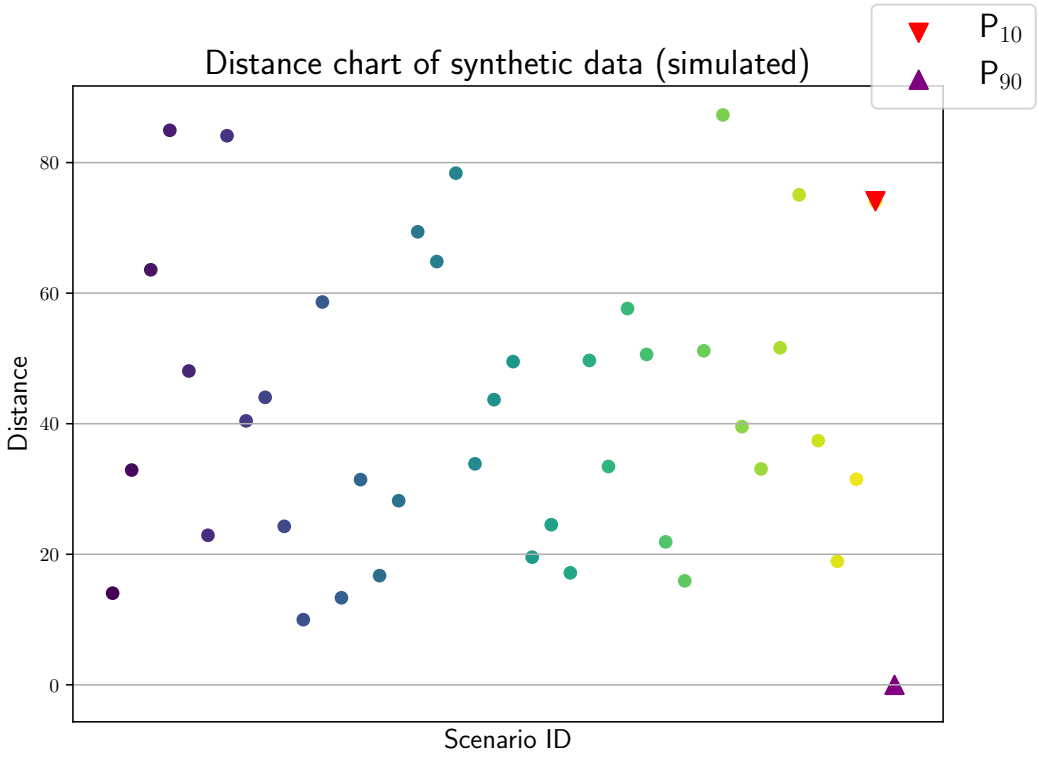



Figure 3.4: Distance chart of our example time-series ensemble. Only the simulated values were used to compute the distances. The baseline used is the $P_{90}$ series. The $P_{10}$ series is show as a red upside-down triangle.

In a Distance chart, the $X$ axis is the series' identifier, and the $Y$ axis is the distance between each series and a baseline. Each series is treated as a multidimensional point when calculating the distance. We use the Euclidean distance, instead of a correlation measure or a dynamic time-warping technique, to maintain coherence with the Bump chart and Time-lapsed LAMP chart. Also, it is computationally light to calculate, even though it becomes increasingly unstable as the data dimensionality increases.

## 3.4 Bump chart

The third visualization technique, the Bump chart, was proposed by (Tufte1990) in order to visualize rankings of objects in time, e.g, cyclists' positions at the end of each day of a *Tour de France* edition. Our proposal differs from the original approach in two main aspects: first, we graph a distance-based ranking built from an ensemble of time-series and a reference time-series; second, the ranking on time step $T$ is calculated by taking into consideration information from time steps $[0, T-1]$, thus, making it a cumulative measure.

A ranking measure is built by comparing a set of elements as they achieve a goal, e.g. athletes finishing a race. However, when dealing with time-series data, the goal may not be clearly defined. Here we defined the goal as the proximity of the ensemble's series to a baseline time-series, using the Euclidean distance. A rank by time measure may be used to assess the adherence of a scenario to the baseline. Depending on the analysis being made, a more adherent scenario may be desirable. However, it may not necessarily be the closest one, in a raw distance sense.

Figure 3.5 shows an example of Bump chart built from the synthetic time-series ensemble. It presents a scale-independent view of the ensemble compared to the baseline, which is presented in red at the $X$ axis. It can also be modified to simply present an ordering of the time-series values at each time step, dismissing the need for a baseline series and presenting the data more closely to the original values.

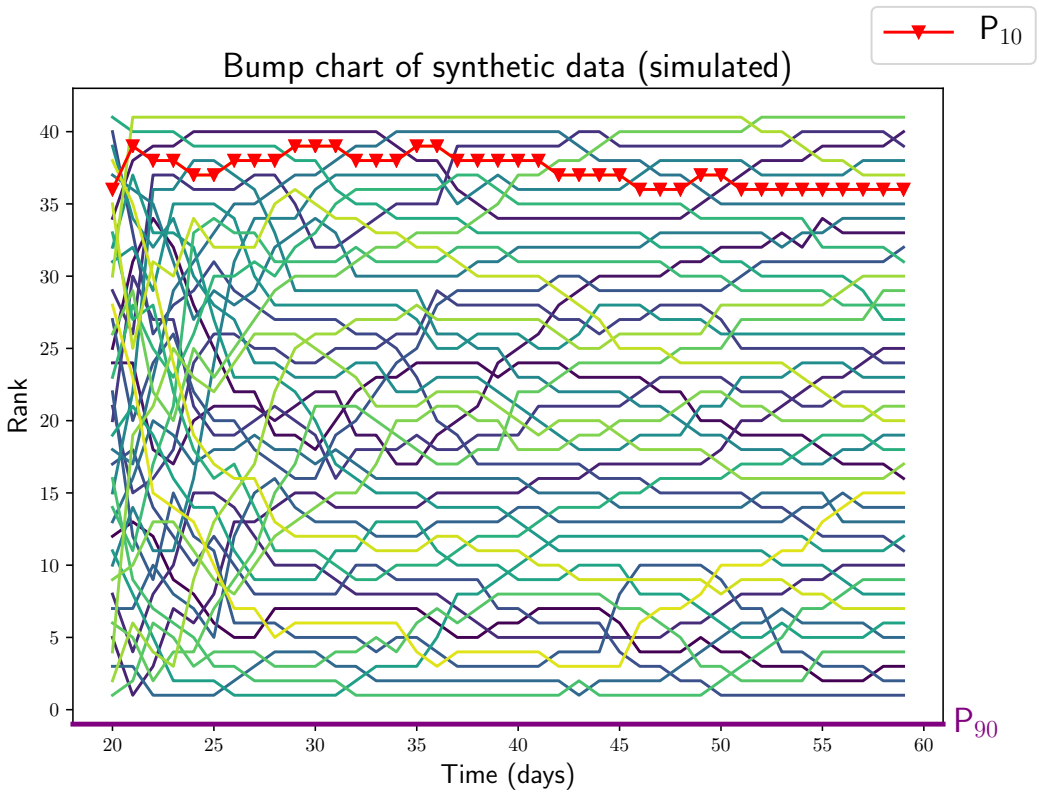



Figure 3.5: Bump chart of our example time-series ensemble. Only the simulated values were used to compute the rankings. The baseline series is indicated in the $X$ axis of the plot, in our case, the baseline is the $P_{90}$ series. The $P_{10}$ series is show in red upside-down triangles.

## 3.5 Local Affine Multidimensional Projection

The Local Affine Multidimensional Projection (LAMP) is a guided multidimensional projection technique that focuses on interactive applications by leveraging the user's knowledge of the dataset's correlations (Joia2011). LAMP draws its basis from orthogonal mapping theory in order to produce accurate mappings given a small subset of samples positioned in the projection space. The main point of LAMP is its interactivity. It allows the user to change the mapping by manually repositioning the samples and then projecting the remaining points again. Figure 3.6 shows the workflow of visual exploration using LAMP.


Figure 3.6: LAMP exploration workflow. Figure originally appeared in (Joia2011).

Differently from Multidimensional Scaling (MDS), LAMP is also classified as a local method. Such methods require neighborhood information for each data instance in order to perform the projection. It also gives LAMP is main advantage: it needs few samples positioned in the projection space to project the remaining points. The locality of LAMP can be leveraged to explore naturally occurring groups in the data (Joia2011). In Subsection 3.5.1 we provide the mathematical definition of LAMP, as stated by (Joia2011), as well as a its algorithm in pseudocode.



### 3.5.1 LAMP definition

Let $X \in \mathbb{R}^m$ be a dataset in its original space, $x \in X$ is a single data instance. The control points set $X_s = \{x_1, x_2, x_3, \ldots, x_k\}$ is a subset of $X$ chosen as an anchor for the mapping process and their correspondence in the projection plane is the set of points $Y_s = \{y_1, y_2, y_3, \ldots, y_k\} \in \mathbb{R}^2$.

The process of building the mapping $x : \mathbb{R}^m \to \mathbb{R}^2$ involves finding an affine transformation that minimizes the sum of weighted differences between the mapping of $x$ and the control points' mappings in $Y_s$. Equation 3-1 presents the mathematical definition of this problem.

$$\begin{aligned} &\text{minimize} && \textstyle\sum_i \alpha_i ||f_x(x_i) - y_i||^2 \\ &\text{subject to} && M^T M = I \end{aligned} \tag{3-1}$$

where the matrix $M$ is an unknown, $f_x(p) = pM + t$ is the unknown affine transformation and $\alpha_i$ are scalar weights defined as the inverse of the squared euclidean distance between $x$ and the points $x_i \in X_s$. Equation 3-2 shows the mathematical definition of these weights.

$$\alpha_i = \frac{1}{||x - x_i||^2} \tag{3-2}$$

The restriction $M^T M = I$ imposed in the minimization problem presented in 3-1 ensures that the transformation behaves like a rigid transformation, meaning it avoids scaling and shearing effects and preserves the original distances as well as possible, even if the control points mapping, which is defined by the user, introduces unavoidable errors in the process.

If we take the partial derivatives of $f_x$, with $t = 0$, we write $t$ in terms of $M$, as in Equation 3-3.

$$\begin{aligned} t &= \tilde{y} - \tilde{x}M \\ \tilde{x} &= \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} \\ \tilde{y} &= \frac{\sum_i \alpha_i y_i}{\sum_i \alpha_i} \end{aligned} \tag{3-3}$$

Given the results in Equation 3-3, the minimization problem in Equation 3-1 can be rewritten as follows:

$$\begin{aligned} &\text{minimize} \quad \textstyle\sum_i \alpha_i ||\hat{x}_i M - \hat{y}_i||^2 \\ &\text{subject to} \quad M^T M = I \end{aligned} \tag{3-4}$$

where $\hat{x}_i = x_i - \tilde{x}$, and $\hat{y}_i = y_i - \tilde{y}$. The minimization problem stated in Equation 3-4 can be rewritten in matrix notation, as shown in Equation 3-5.

$$\begin{aligned} &\text{minimize} \quad ||AM - B||_F \\ &\text{subject to} \quad M^T M = I \end{aligned} \tag{3-5}$$

where $||.||_F$ is the Frobenius norm, and matrices $A$ and $B$ are defined as in Equation 3-6.

$$A = \begin{bmatrix} \sqrt{\alpha_1}\hat{x}_1 \\ \sqrt{\alpha_2}\hat{x}_2 \\ \vdots \\ \sqrt{\alpha_k}\hat{x}_k \end{bmatrix}, B = \begin{bmatrix} \sqrt{\alpha_1}\hat{y}_1 \\ \sqrt{\alpha_2}\hat{y}_2 \\ \vdots \\ \sqrt{\alpha_k}\hat{y}_k \end{bmatrix} \tag{3-6}$$

The minimization problem rewritten in Equation 3-5 is an instance of the Orthogonal Procrustes Problem (Gower2004), which has a known solution presented in Equation 3-7.

$$M = UV, A^T B = UDV \tag{3-7}$$

where $UDV$ is the Singular Value Decomposition of $A^T B$. With the known value of $M$, the value of function $y = f_x(x) = (x - \tilde{x})M + \tilde{y}$ can be calculated, thus, resulting in the projection $y \in \mathbb{R}^2$ of data instance $x$. Algorithm 3.5.1 shows a pseudocode of LAMP.

One of the advantages of LAMP over other MP techniques is its axis stability. Unlike MDS techniques, LAMP does not suffer from axis rotations, so it does not require post-processing the maps to conform to the same

**Algorithm 3.5.1:** LAMP algorithm

**Data:** set $X \in \mathbb{R}^m$, set $X_s \in X$ of control points and set $Y_s \in \mathbb{R}^2$ of mappings of the control points.
**Result:** set $Y \in \mathbb{R}^2$ of mappings of $X$.

**for** $x \in X$ **do**
  $\alpha \leftarrow \mathbf{0}$
  **for** $x_i \in X_s$ **do**
    $\alpha_i \leftarrow \frac{1}{||x_i - x||^2}$
  **end**
  $\tilde{x} \leftarrow \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i}$
  $\tilde{y} \leftarrow \frac{\sum_i \alpha_i y_i}{\sum_i \alpha_i}$
  $\hat{x}, \hat{y} \leftarrow \mathbf{0}$
  **for** $\hat{x}_i \in \hat{x}$, $\hat{y}_i \in \hat{y}$ *and* $y_i \in Y_s$ **do**
    $\hat{x}_i \leftarrow \tilde{x} - x$
    $\hat{y}_i \leftarrow \tilde{y} - y_i$
  **end**
  $A, B \leftarrow \mathbf{0}$
  **for** $\alpha_i \in \alpha$, $\hat{x}_i \in \hat{x}$ *and* $\hat{y}_i \in \hat{y}$ **do**
    $A_i \leftarrow \sqrt{\alpha_i}\hat{x}_i$
    $B_i \leftarrow \sqrt{\alpha_i}\hat{y}_i$
  **end**
  $U, D, V \leftarrow \mathrm{SVD}(A^T B)$
  $M \leftarrow UV$
  $y \leftarrow (x - \tilde{x})M + \tilde{y}$
**end**

orientation. This feature is the main reason behind the choice of LAMP as the basis of our proposal.

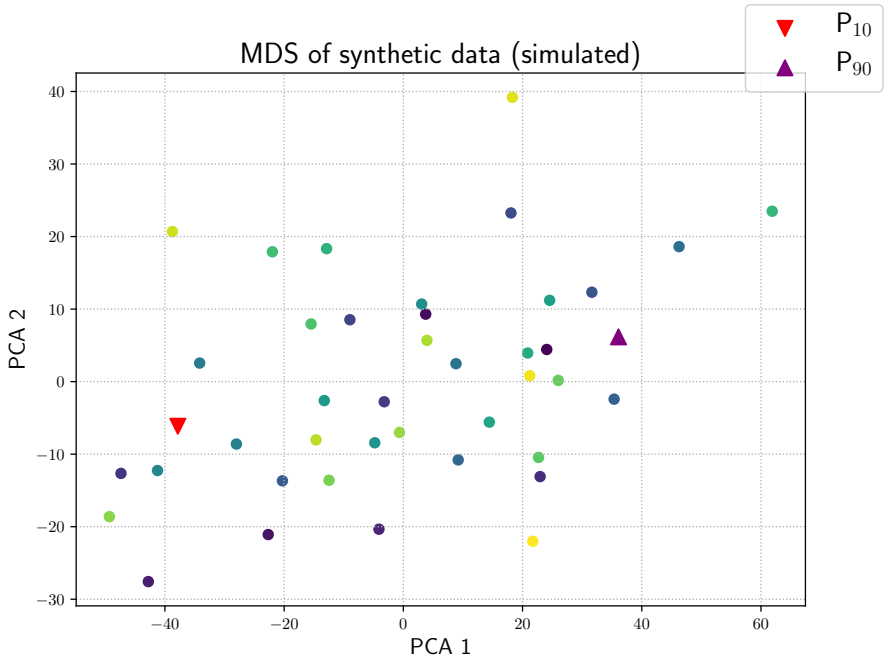
## 3.6 Time-lapsed Local Affine Multidimensional Projection

Building on the concepts presented in Section 3.5, our final view is a modified version of LAMP, where each time-series in our ensemble is projected as a series of points at each time, thus presenting its evolution compared to all other series in the ensemble.

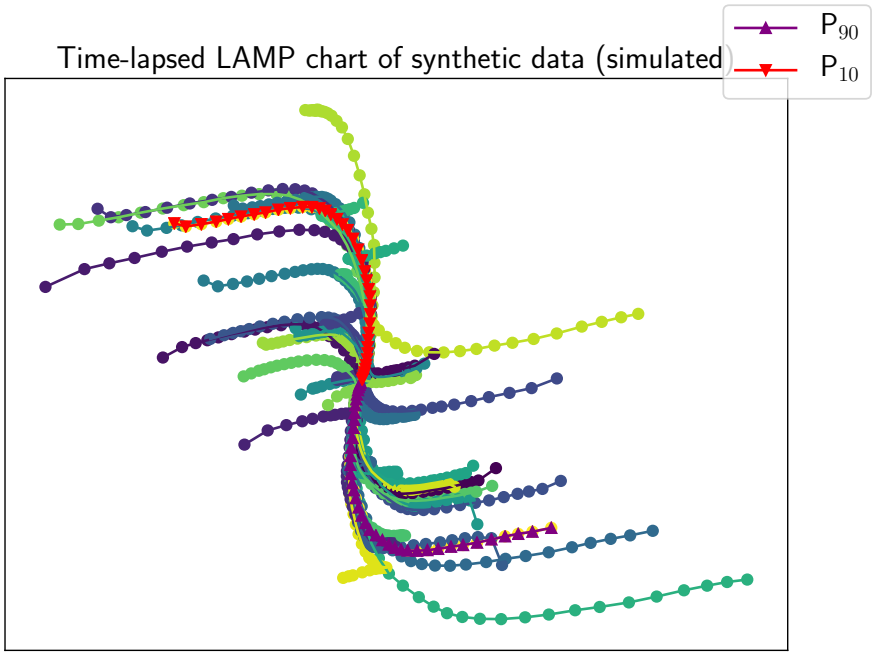
Given an ensemble $S$ of time series with $T$ time steps each, we build $T-1$ mappings. Each mapping $t \in [2, T]$ is composed by the projected data from times $[1, t]$, and is independent from the others. However, when merged in a single view, the result is a series of $|S|$ paths, as shown in Figure 3.7b. Each path traces the evolution of a single time series, allowing its comparison to the $|S| - 1$ others.

This approach can be used to find features which are not easily verifiable

using ordinary MP approaches. Figure 3.7 shows an example of four synthetic curves projected using MDS (3.7a) and our Time-lapsed LAMP (3.7b).




(a) MDS projection of our ensemble using only simulated data.




(b) Time-lapsed LAMP chart of our ensemble using only simulated data.

Figure 3.7: Comparison between the MDS projection and Time-lapsed LAMP projections of our example ensemble.

The original LAMP algorithm requires three parameters: $X \in \mathbb{R}^T$ as the data to be mapped, $X_s \in X$ as the control points, and $Y_s \in \mathbb{R}^2$ as the mapping of $X_s$ onto the projected space. The first step in our approach is to build

the control points set $X_s$. We opted to use the whole set $S$ as control points; therefore, they must be positioned in the projection space before the main mapping step. For this task, we employed an MDS algorithm (Kruskal1978), using the $|S|$ series and $T$ time steps as input. We also opted to use an Euclidean distance between each time series as metric for this step. The resulting projected points are then used as the $Y_s$ parameter for LAMP, while the original series $S$ is used as the $X_s$ parameter. The $X$ parameter, however, needs special processing. Since $X$ constitutes the data to be mapped, and we build one mapping for each time step $t$, $X$ must be adapted to contain only the data to be projected up to time step $t$. This adaptation is done by replacing the data of $S$ outside the time step range $[1, t]$ with zeros; this effectively removes that range from consideration for the mapping. An outline of the procedure is as follows:

1. Build the control points mappings $Y_s$;
   - Use an MDS algorithm with the $T$ time steps of all $S$ time series;
2. For each time step $t \in [2, T]$, map the time series using their whole data as the $X_s$ parameter and their mappings as the $Y_s$ parameter;
   - To build $X$, use only the values in the time step range $[1, t]$;
   - Replace the remaining values $[t + 1, T]$ with zeros.
3. Merge the resulting mappings and plot them.

Contrary to most MP algorithms, this approach results in a sequence of mappings for each time series. Each mapping presents the behavior of that single series, allowing to compare it to all others in the set. This feature may be useful to detect patterns in the behaviors of sets of scenarios, so that the user may select them for further analysis.

## 3.7 Interaction Mechanisms and Prototype Details

To evaluate the effectiveness of our approach, we developed a prototype application[1] implementing the views presented in this section as well as the brushing and linking technique to add interactivity to our prototype. We implemented the prototype using the Python programming language (version 3)[2]. We used Matplotlib 2.0 (Hunter2007) as a graphical plot library, Qt 5.7[3] as the user interface API, NumPy (Numpy2011), Scipy (Scipy2001) and

[1] The source code is freely available at: `https://tinyurl.com/proto-vs`
[2] Available at: `https://www.python.org`
[3] Available at: `https://www.qt.com`

Scikit Learn (ScikitLearn2012) for the high performance, numerically heavy computations.

In order to show the interaction process developed, we included several figures with examples below. Full resolution versions of these figures are also available online on the source-code repository, as well as a video showing the interaction process. Figure 3.8 presents the initial window of our prototype after the ensemble is loaded.

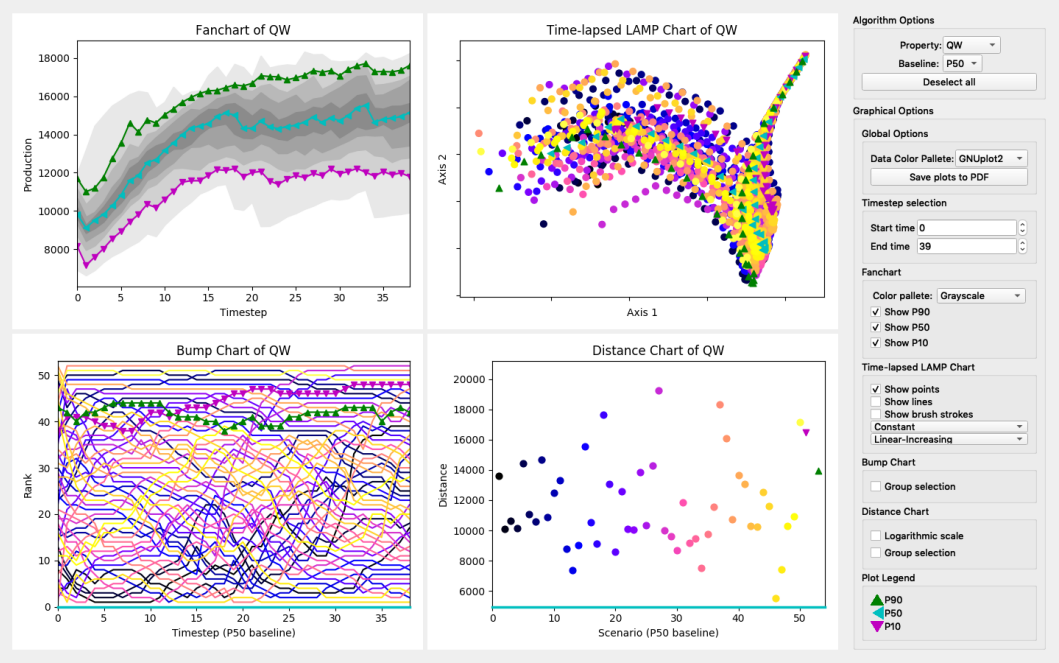
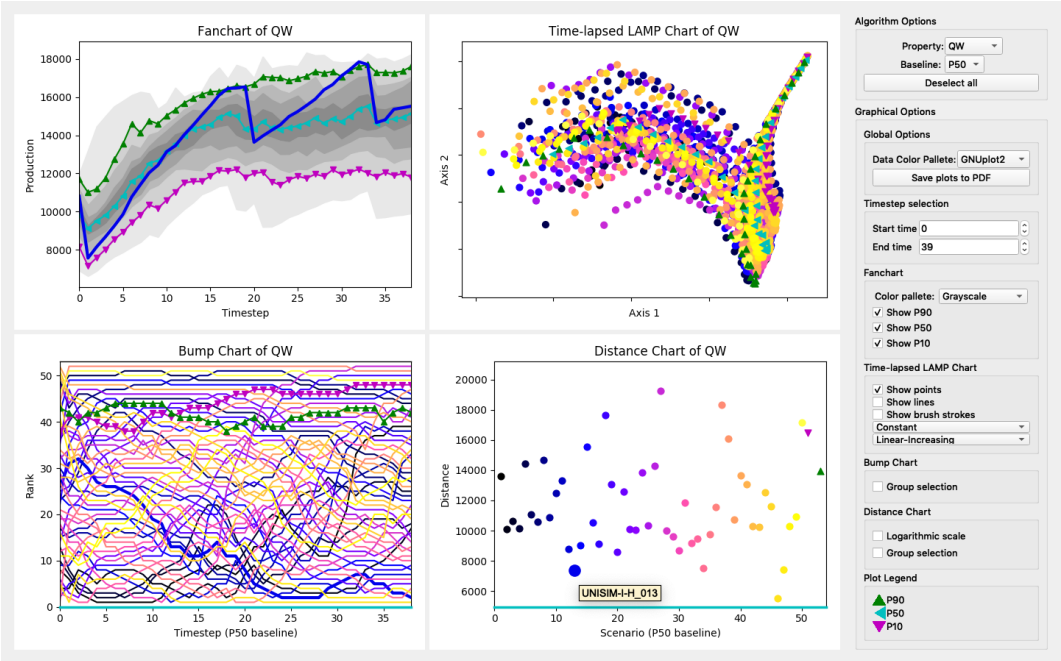
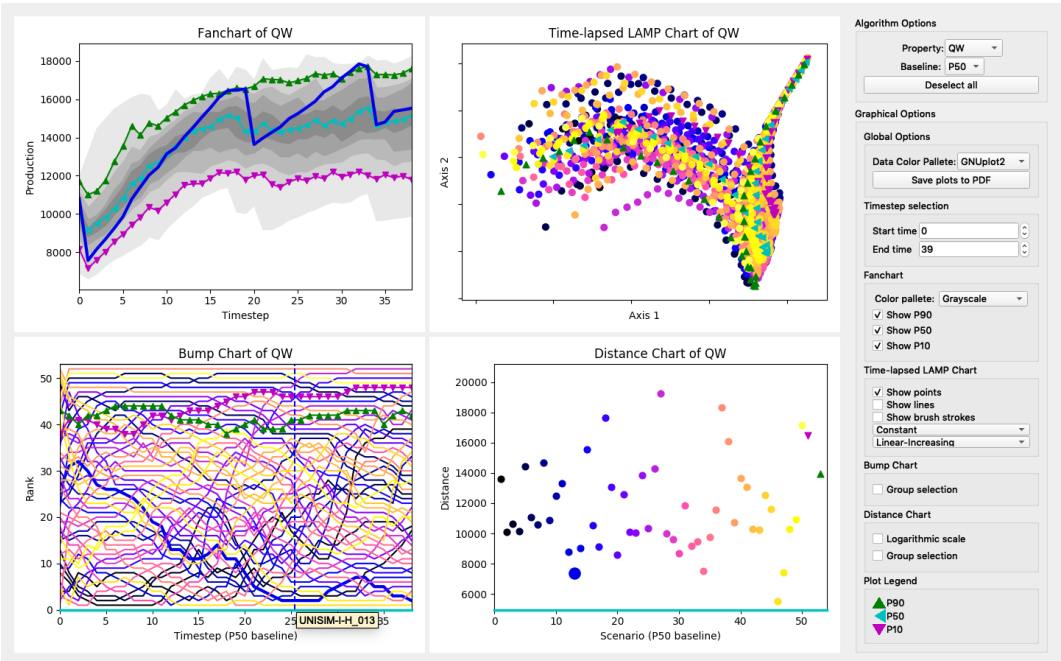



Figure 3.8: Initial window of our prototype after an ensemble is loaded.

Most of the application window is covered by the charts explained in sections 3.2, 3.3, 3.4 and 3.6. To the right, there is a control panel with several options regarding the ensemble, selection and graphical options as well. Figure 3.9 shows the mouse-over highlighting interaction mechanism. This feature is activated when the user positions the mouse over a scenario in any of the charts, except the Fanchart. The graphical representation of that scenario is highlighted in all charts, and its name is shown as a tooltip near the mouse cursor.




(a) Mouse-over interaction on the Distance Chart. Note the tooltip over highlighted on the Distance Chart.

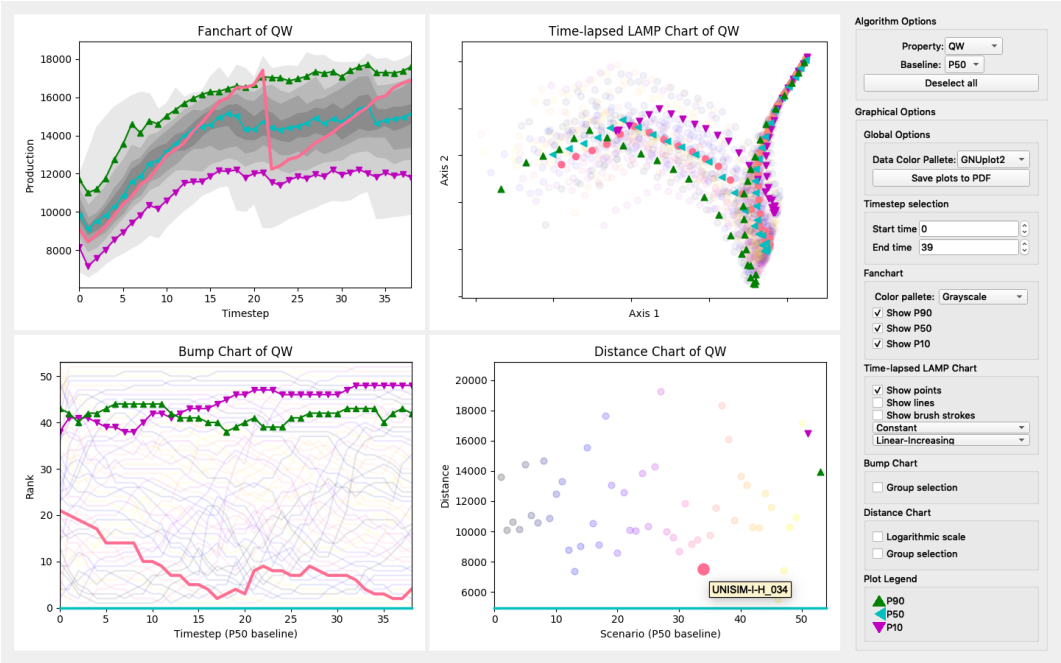



(b) Mouse-over interaction on the Bump Chart. Note the tooltip over highlighted on the Bump Chart.

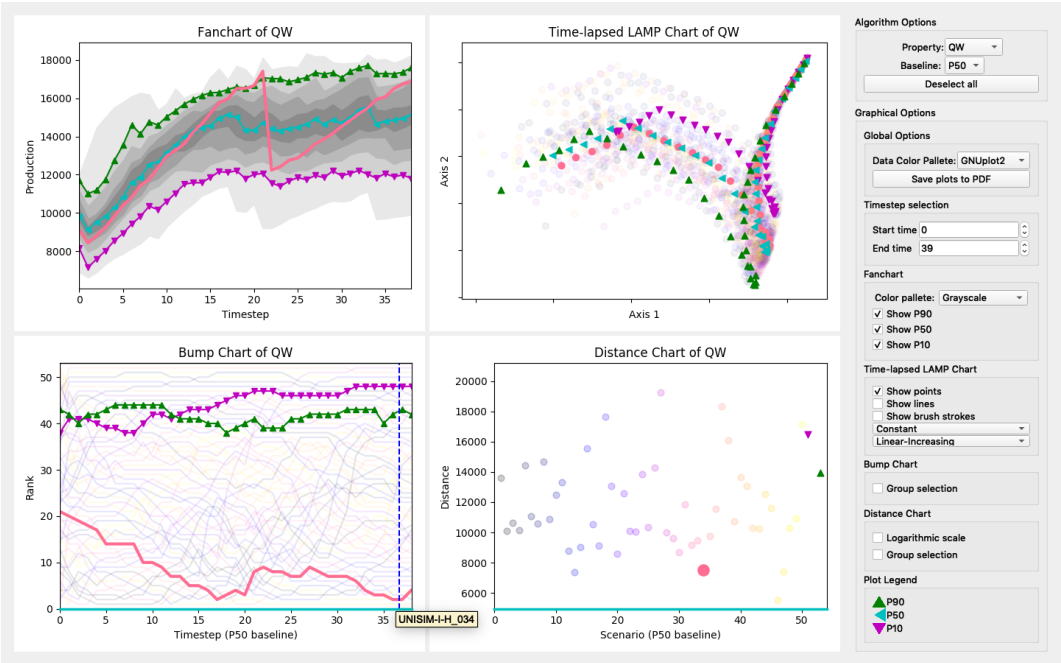
Figure 3.9: Example of the mouse-over interaction on the Distance Chart (3.9a) and the Bump Chart (3.9b).

In order to select a scenario, the user can click in its representation in any of the charts. By doing so, that scenario is highlighted in all charts by means of reducing the other scenarios opacity to a minimum value, as shown in Figure 3.10. Multiple scenarios can be highlighted in this manner. Clicking

on a highlighted scenario will make it transparent again. When the last scenario is de-selected, all scenarios will become opaque, as in Figure 3.8.




(a) Selecting a single scenario on the Distance Chart. Note the tooltip over the selected scenario on the Distance Chart.

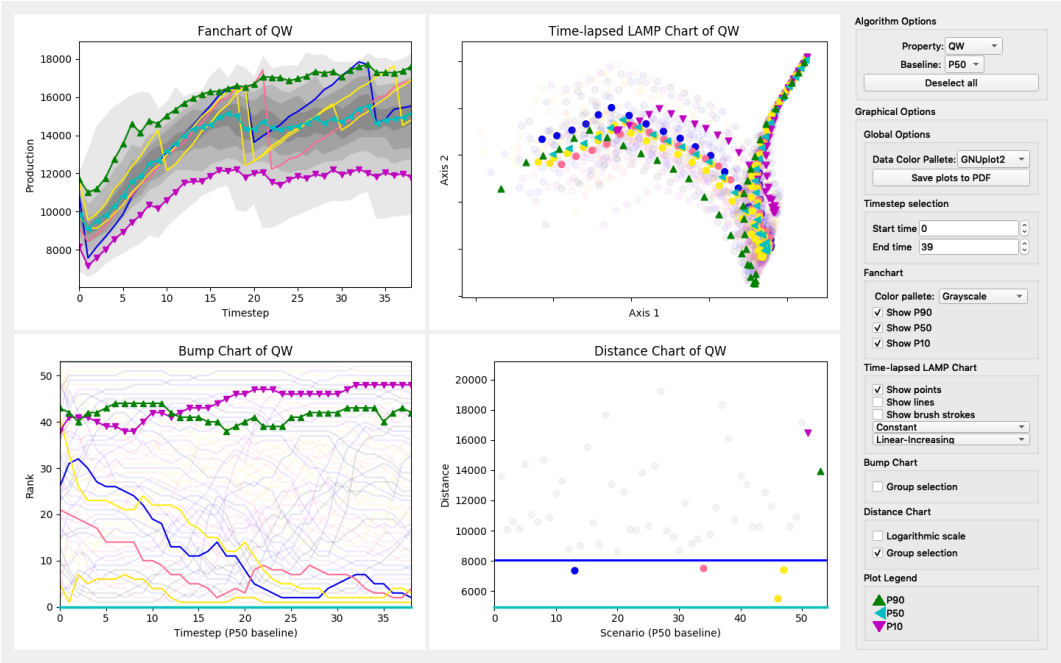



(b) Selecting a single scenario on the Bump Chart. Note the tooltip over the selected scenario on the Bump Chart.

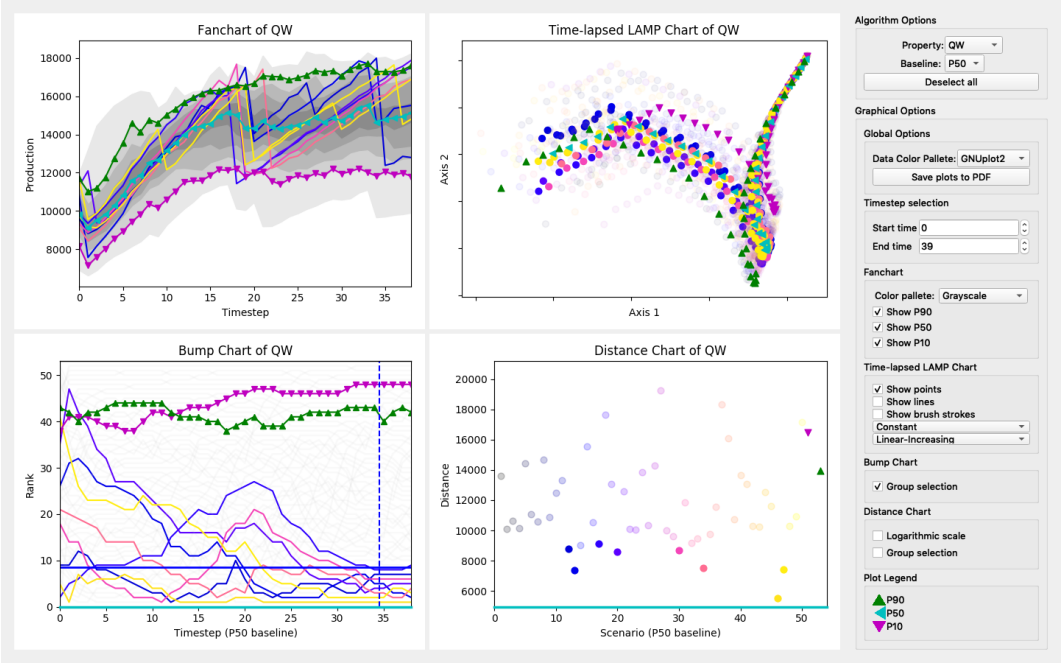
Figure 3.10: Example of the highlight interaction on the Distance Chart (3.10a) and the Bump Chart (3.10b).

Besides the click-to-highlight mechanism, the Bump and Distance charts also provide a group selection feature. In the Bump chart, all scenarios with a

rank better than the threshold (defined by the $Y$ value of the mouse cursor) at the current time (current $X$ value) will be selected if the user clicks the left mouse button. A similar logic is implemented for the Distance chart; all scenarios with a distance lesser than the one pointed by the mouse cursor are selected if the user clicks with the left mouse button. Figure 3.11 shows an example of the group selection feature for both charts mentioned before.




(a) Highlighting a group of scenarios on the Distance Chart.




(b) Highlighting a group of scenarios on the Bump Chart.

Figure 3.11: Example of the highlight interaction being used to select a group of scenarios on the Distance Chart (3.11a) and the Bump Chart (3.11b). For the Distance Chart, all scenarios with a distance lower than an user-selected threshold are highlighted, while for the Bump Chart, only the scenarios with a rank better than an user-selected threshold at a given time are selected.

The middle mouse button performs pan operations on all charts when clicked. When the mouse-wheel is scrolled a zoom operation is performed. These features are shown for the Fanchart, in Figure 3.12. However, these actions can be done on all charts.

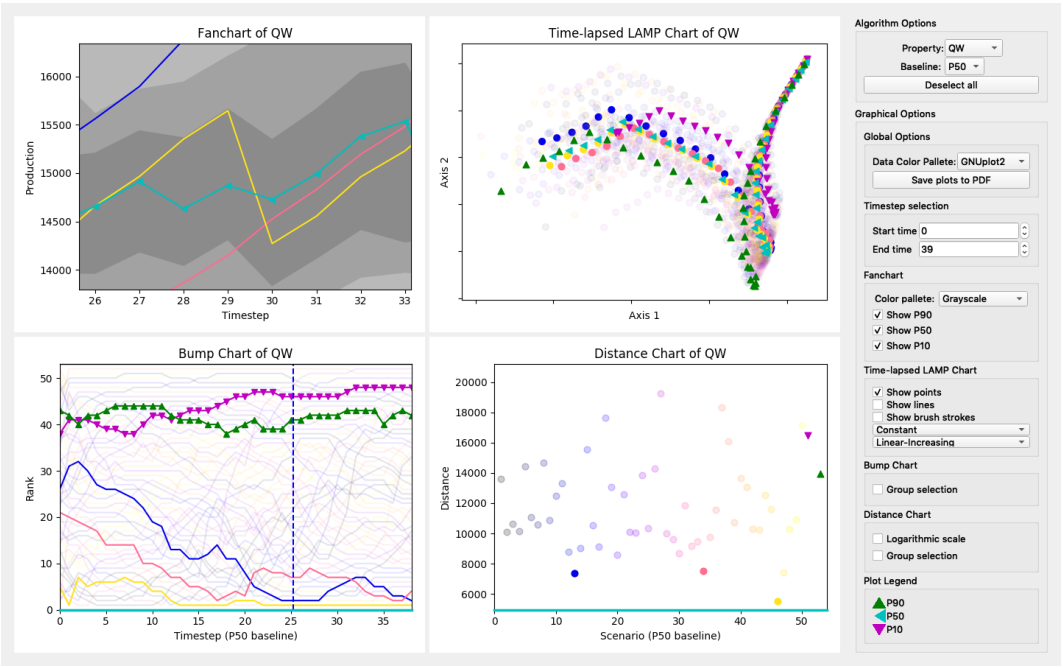



Figure 3.12: Pan and zoom actions performed on the Fanchart.

# 4
# Evaluation and Results

Our proposal is a visual interactive approach for the scenario reduction problem. In short, we propose a way to graphically explore the results of a series of simulations and visually perform a scenario reduction task and assess its results.

In order to test our proposal, we developed a prototype application and conducted an empirical evaluation study with potential users. We also compared the user-selected scenarios to the results of other scenario reduction approaches found in the literature and industry.

This chapter is organized as follows: Section 4.1 describes the data we used for the tests; Section 4.2 describes the empirical study; Section 4.3 presents the results of our study and some discussions on their implications; Section 4.4 presents some improvements we made based on our study; finally, Section 4.5 compares the scenarios selected using our approach and other approaches found both in the literature and in industry.



## 4.1
## Test ensemble

The ensemble we used in our tests comes from a synthetic model called UNISIM-I (Avansi2015), created for testing algorithms and methodologies related to reservoir management. This model was built using real publicly available data from the Namorado Field located in the Campos Basin, Brazil. It comprises high-quality geological and production data to ensure that any derived models honor the original data (Avansi2015). The base model contains a set of four exploratory wells used to estimate the initial values of the reservoir's production and petrophysical properties. Based on this initial model, a production strategy was defined by adding a number of wells and preparing an ensemble of 200 realizations for reservoir simulation using the IMEX simulator. The resulting simulations have a high degree of uncertainty, which was reduced by performing a history matching process using an ensemble-based method (ensemble smoother with multiple data assimilation (Emerick2013)). The history matching considered oil and water production rates ($Q_o$ and $Q_w$, respectively), gas-oil ratio at producing wells, and bottom-hole pressure at production and

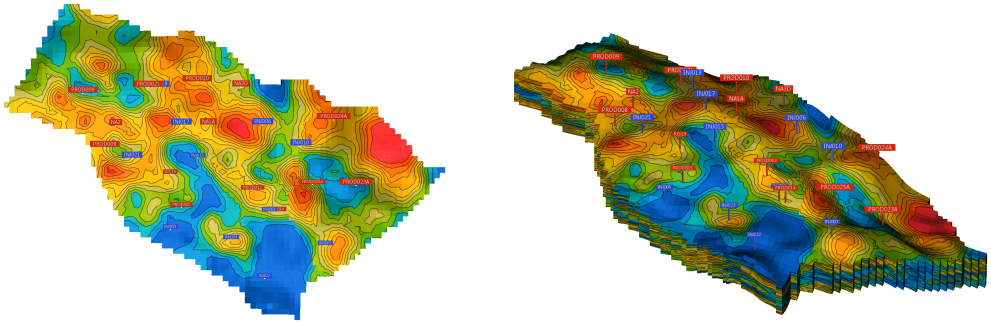


Figure 4.1: UNISIM-I-H geometry with the producer wells in red and injector wells in blue. The grid property shown is the field porosity.

water injection wells, for a period of 10 years. The uncertainty parameters correspond to porosity, net-to-gross ratio, horizontal and vertical permeability at every reservoir gridblock, end-points of water relative permeability curve, rock compressibility, and water-oil contact. This resulted in another ensemble with lower uncertainty, thus more appropriate for production forecasting.

The resulting simulations contain a set of 25 wells, the 4 exploratory ones, plus 21 added by the engineers who defined the production strategy. Each well can be classified as either injector or producer. The models used in our tests are composed of 14 producer and 11 injector wells. The focus of our analysis lied on the cumulative oil and water production wells ($N_p$ and $W_p$, respectively). Each simulated model contains 30 years of data: 10 years of observed data and 20 years of production forecasts; the data is sampled monthly for the historic data, and every 6 months for the forecasts. Figure 4.1 shows the reservoir geometry and the location of the selected wells.[1]

For the study, we imported oil and water cumulative production data, as well as their production rates of the producer wells, saving data from each well and property in a comma-separated-values (CSV) file. For each scenario, we summed the productions of all wells, in order to obtain a single time series for each property of each scenario. The first 10 years of historical data were removed, since the goal was to perform the analysis using only forecast production data.

## 4.2 Evaluation Method

To evaluate our proposal, we performed an observational study according to the following procedure: Each participant was given an overview of the

[1]Images produced using Geresim: `http://webserver2.tecgraf.puc-rio.br/~celes/projects.html`

prototype and its features, and three independent tasks related to scenario reduction. The evaluators observed and recorded the participants interacting with the prototype to perform the proposed tasks. Throughout the session, the evaluators asked participants to give feedback on the prototype, aiming to identify interaction flaws and opportunities for improvement.

The study was divided in three rounds. At each round, the researchers collected and prioritized the identified problems and suggestions, and made the corresponding corrections and improvements in a new version of the prototype, incrementally improving the user-system interaction before a final evaluation round, which involved experts in scenario reduction.

Before each session, the researchers assessed the participants' knowledge by means of a questionnaire, which is available at Annex A[2]. The questionnaire asked about the users' background and professional areas, as well as specific knowledge about the graphical views, the definition of a percentile series, and time-series analysis. Answers to these questions were ranked on a five-point scale, ranging from 1 (no knowledge) to 5 (expert knowledge). Afterwards, we asked for the participant's consent of having their voice and interaction recorded during the interview. The participants opted to give consent or not, and signed a form acknowledging their decision. Should a participant forbid the recordings, we terminated the interview, and any data collected, including their answers to the pre-session questionnaire, were deleted from our records.



## 4.3 User study Results

A total of 29 people participated in the study: 11 on the first round, 8 on the second, and 10 on the third round.

### 4.3.1 First round

The first round was considered a preliminary study, and was conducted in order to assess the feasibility of our approach for scenario reduction and to help plan a more detailed study. The eleven participants of this study are all laboratory colleagues of the author, all of them with science and technology backgrounds, comprising eight graduate and three undergraduate students, with varying levels of education: one D.Sc., six M.Sc., one B.Sc., and three undergraduate students in Electronics Engineering. None of them had prior knowledge of the scenario reduction problem before the evaluation sessions.

[2]The questionnaire is also available online at `https://tinyurl.com/vs-reduction`.

We asked participants to find 3 to 5 scenarios closest to the $P_{50}$ scenario calculated from the ensemble, using the Cumulative Water Production ($W_p$) property. Most users had no problems in finding a set of scenarios. Most of the issues raised by them were related to the user-system interaction. Some users commented on the cluttering of the Bump chart, which is a problem the researchers had somewhat anticipated during the development phase of the prototype. They also reported some confusion regarding the Time-lapsed LAMP chart. Most users found it hard to interpret, since it has no direct connection to the original data, but it is rather a representation of the distances between the scenarios at each time step.

### 4.3.2 Second round

The second evaluation round was composed exclusively of undergraduate students with no prior knowledge of the scenario reduction problem. These students were recruited from Human-Computer Interaction classes of the Department of Informatics of PUC-Rio. All participants of this round work on information technology related areas and have a varying level of experience, ranging from the first to the last semester in their courses.

For this round, we opted for participants who were unlikely to know the target domain. This allowed us to assess the difficulty of the task in its worst case, as well as any user interface issues that could interfere with the study. Since little or no previous knowledge could be an obstacle to the study, we prepared a small tutorial in order to explain the concept of percentiles, and to present the prototype, its graphical views and interaction mechanisms. This tutorial was given after the pre-session questionnaire. The participants were also given an opportunity to use the prototype at will for a few minutes, in order to clarify any doubts about the concepts presented in the tutorial. We noticed that, since these participants were not used to this area, they paid more attention to the interaction mechanisms, and made several comments that helped us refine the prototype even further.

In this round, regarding the graphical views, two participants claimed to have moderate knowledge about the distance chart and one participant claimed to have moderated knowledge about the fanchart. As for the other visualizations all participants claimed to have little or no knowledge about them. Regarding the concept of time-series analysis, one participant claimed moderate knowledge, while all others claimed to have little to no knowledge of it. However, the proportion of users that had seen the concept of percentiles or time-series analysis is higher compared to the proportion of users claimed

some knowledge about the proposed graphical views. Figure 4.2 shows their answers to the questionnaire.

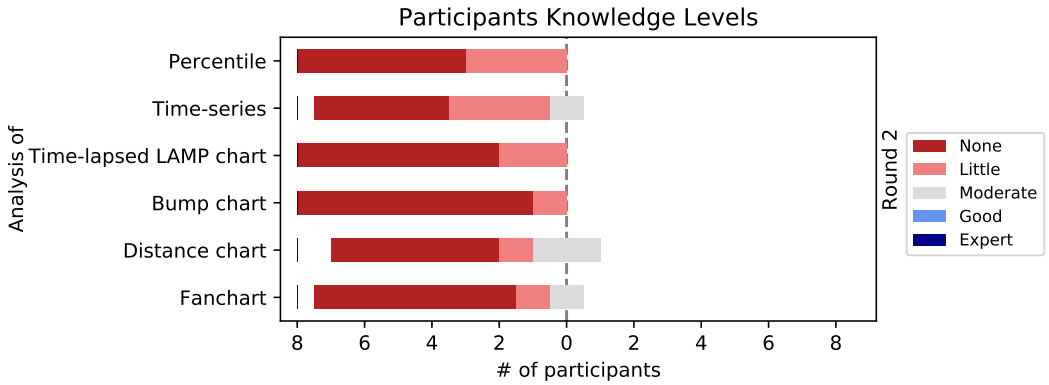



Figure 4.2: Second round participants' answers to the knowledge-based questions of the pre-session questionnaire.

After the first four evaluation sessions for this round, we noticed that participants were confused by the bump chart's ranking concept. They seemed to associate a higher numerical rank with better adherence to the reference scenario, which was not the case. In order to further evaluate this issue, the bump chart was changed for the last four sessions, in order to present a better adherence with a higher numerical rank. Some participants understood this ranking concept more naturally after the change. However, more studies must be performed in order to investigate whether previous knowledge is necessary to properly interpret this chart.

The main issues found during this evaluation round were: (i) confusion between rank and score with the Bump chart; and (ii) confusion between time steps and scenario IDs with the Distance chart.

And the main improvements suggested by the participants in the second round were: (i) "undo" option for the selection; (ii) band-based group selection for the Bump and/or Distance charts; (iii) filter out the time steps of the Fan and Bump chart using an X-axis zoom feature; (iv) filter out scenarios using the Distance chart group selection feature (remove the scenarios above the threshold line); and (v) adopt similar behavior for the Bump Chart.

Before the experiment started, we expected the users to rely heavily on the Distance chart, because the task was to find the scenarios closest to a pre-defined baseline. Such task naturally invites the user to rely upon the Distance chart, while using the other views as auxiliary aids for the selection. We also expected some confusion regarding the Bump and Time-lapsed LAMP charts, since the information presented by them is much denser compared to the other charts. However, most users of the second round used the Bump and

Distance charts as their main drivers, while the Fanchart was used more as a reference, which partly conforms to our expectations. Also, some users did not use the selection mechanism provided, relying only on the tooltips and highlights between the views. The users who did use the selection mechanism, used the group selection to find an initial set of answers, and then analyzed those in order to make their choice, which is the expected behavior (*i.e.*, the behavior we designed for).

### 4.3.3 Third round

For the third evaluation round, the participants were 8 professors (with PhD degrees), 1 Post-doctoral researcher working in different departments of PUC-Rio, and 1 Software Engineer working outside the University. They all work with – or have extensive academic experience on – scenario reduction and its applications (see Figure 4.3). As expected, most of them claimed to have good knowledge or expertise in the percentile and time-series analysis areas, *i.e.*, their knowledge of the related subjects was much higher, compared to the participants of the second round. Therefore, the tutorial made for the second round participants was not administered to these users.

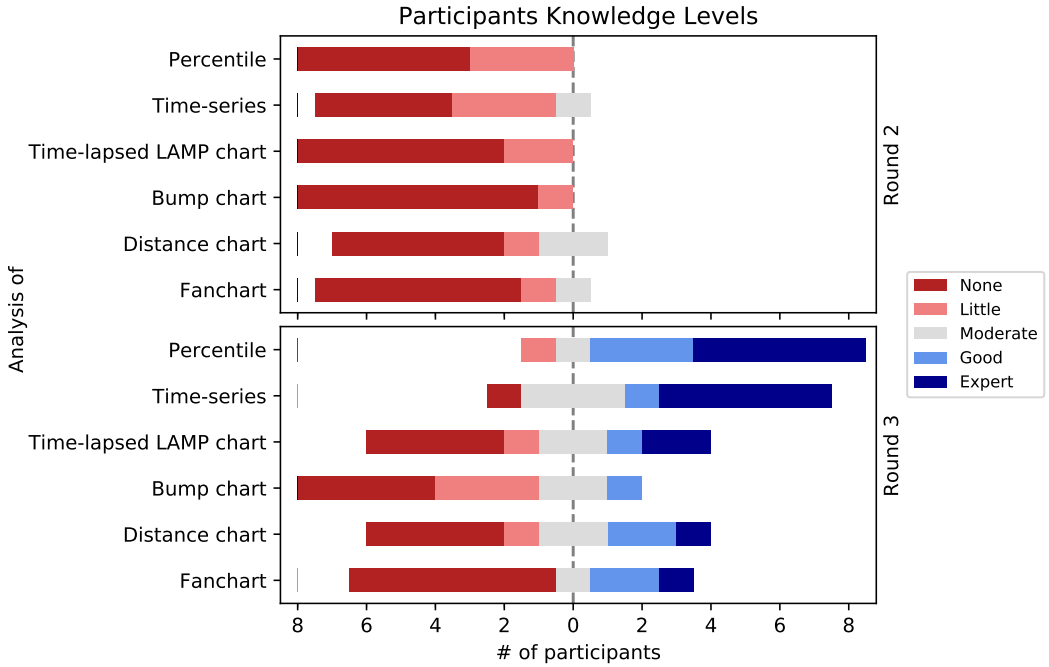



Figure 4.3: Second and third round participants' answers to the knowledge-based questions of the pre-session questionnaire.

However, their knowledge of the graphical views was poorer compared to the percentile and time-series analysis concepts. While for the concept of

percentiles five users claimed to be experts and three users claimed to have good knowledge, for the graphical views, at least four users claimed to have no knowledge of each view. For the fanchart, for instance, six users claimed have no knowledge of it. We gave the participants an overview of the prototype and its features, but an explanation about percentiles, time-series, and scenario reduction was not necessary. Of the 10 participants of this round, only 8 solved all tasks proposed; the other 2 provided an extremely rich discussion but, unfortunately, their available time ran out after more than 2 hours of discussions and little usage of the prototype.

During the observation sessions, the participants did not pay too much attention to the user interface, but rather focused on the scenario reduction task. All participants made extensive use of the mouse hover to highlight features, while only a few used the click-to-select feature.

During the post-task interviews, all participants commented on the difficulty of scenario reduction problems, especially considering that a crucial and potentially costly decision must be made based on the results of this task. They appreciated the possibility of visually exploring the ensemble, mainly because any visual patterns can then be easily identified. When asked about the existence of other visual approaches, the participants claimed to not know any similar approaches to ours for this particular problem. We also asked about other areas besides oil & gas that they thought would benefit from this approach. Four participants mentioned their own research area, energy resource management, since it involves similar data and problems, but instead of oil/water production they deal with solar, wind, and water-based electric energy resources and scenarios where sunlight, wind, and rainfall may not be enough to sustain the energy generation demand during the whole year, thus requiring the more expensive thermoelectric plants to be activated. Another participant mentioned stock management and distribution on big chain-stores involving varying demands for certain items based on region, seasonality, and managing uncertain market conditions. Their feedback gives us confidence on the novelty and general applicability of our work. When asked about the most helpful graphical views, all participants answered either the Bump or Distance charts. However, none of them extracted meaningful information from the Time-lapsed LAMP chart.

## 4.4 Uncertainty and Time Encoding

During the course of the interviews, we noticed that most users had issues with the level of abstraction presented by the Time-lapsed LAMP chart,

namely, they had difficulty telling how the time was visually represented. Thus, we propose the use of explicit encoding (Szafir2018) in order facilitate the perception of time in this particular view. Figure 4.4 shows a comparison between three ways to represent the concept of time in scatterplots: (i) Juxtaposition; (ii) Superposition; (iii) Explicit Encoding (Szafir2018).

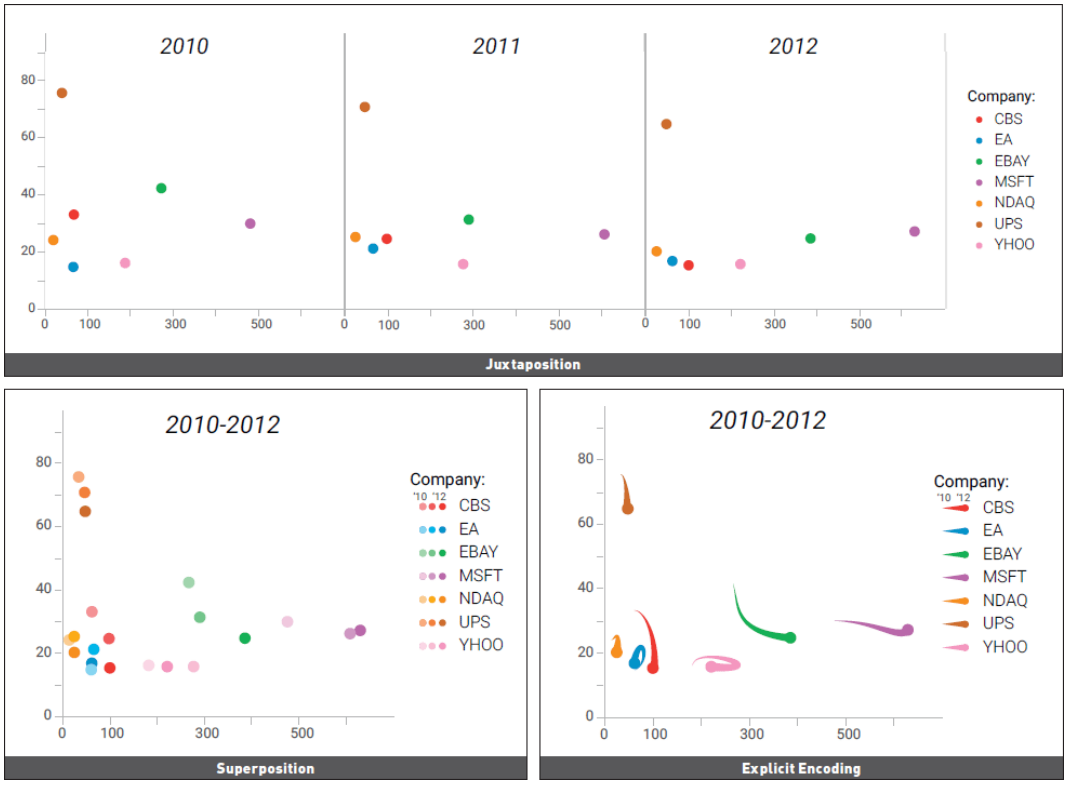
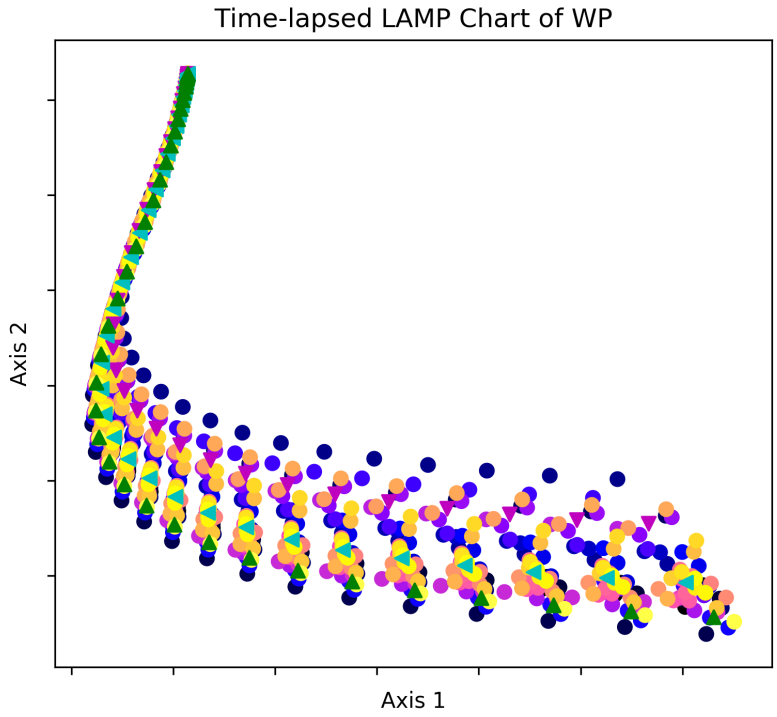
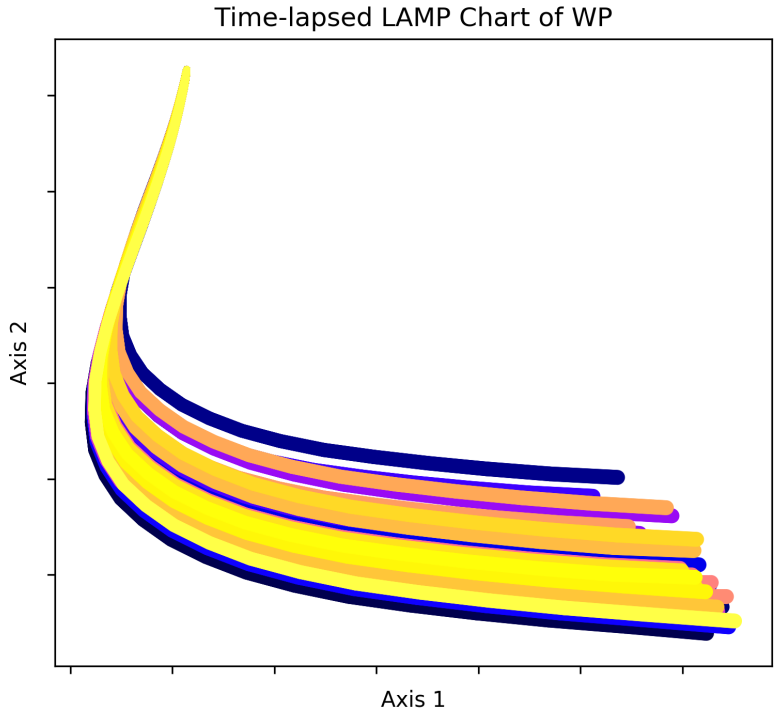



Figure 4.4: Juxtaposition, Superposition and Explicit Encoding. Image taken from the work of (Szafir2018).

In our user-study, we employed the superposition in order to present the projections across time. With the results of the user-study, we propose to use Explicit Encoding to help the users gauge the passing of time, as shown in figures 4.5 and 4.6. For both figures, we use fifty scenarios in order to avoid cluttering the Time-lapsed LAMP chart.

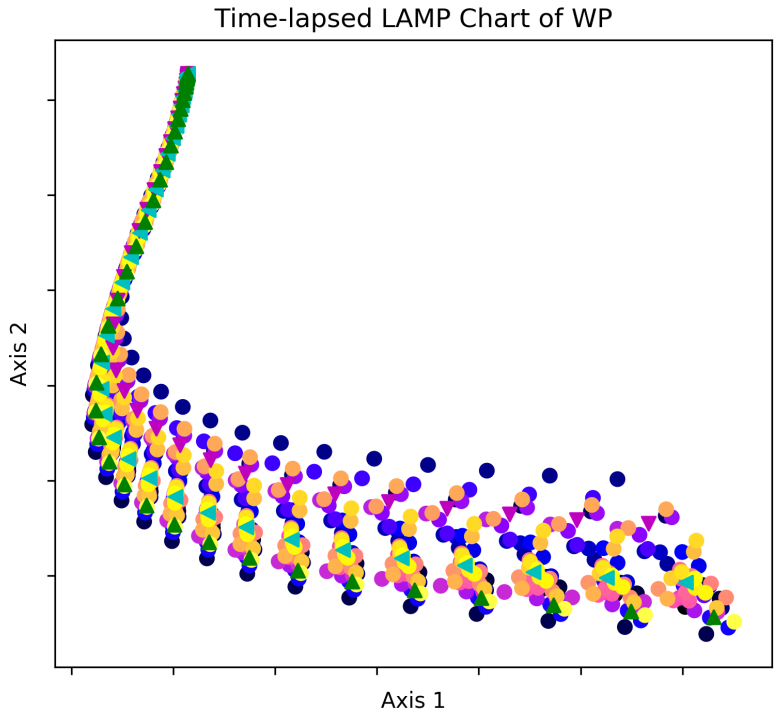



(a) Superposition of times in the Time-lapsed LAMP chart.

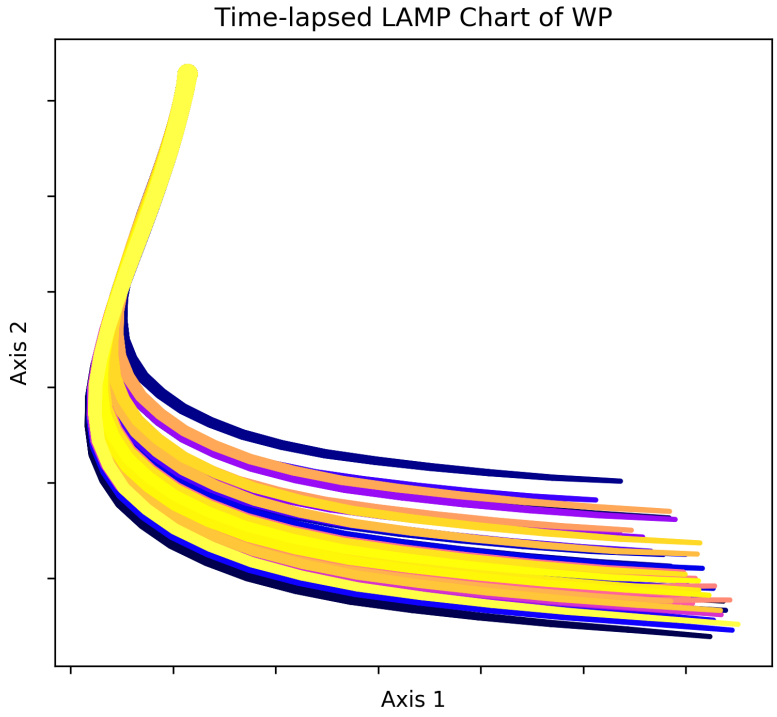



(b) Explicit encoding of time in the Time-lapsed LAMP chart. Starting times encoded with smaller glyphs.

Figure 4.5: Comparison between different time encodings on the Time-lapsed LAMP chart of $W_p$. Superposition (4.5a) and Explicit Encoding (4.5b) from small to large glyphs.




(a) Superposition of times in the Time-lapsed LAMP chart.




(b) Explicit encoding of time in the Time-lapsed LAMP chart. Starting times encoded with larger glyphs.

Figure 4.6: Comparison between different time encodings on the Time-lapsed LAMP chart of $W_p$. Superposition (4.6a) and Explicit Encoding (4.6b) from large to small glyphs.

In figures 4.5 and 4.6 we show two approaches for representing the time:

(i) glyphs presented in ascending order of size (4.5b); (ii) glyphs presented in descending order of size (4.6b). The matter of increasing or decreasing glyph sizes may be a matter of personal preference; however, since the scenarios' projections are close to each other, the encoding's benefits may not be so apparent when using increasing glyph size (4.5b) compared to decreasing size (4.6b).

The glyph size can also be used to represent the uncertainty of the data. However, representing two different variables with a single visual encoding presents visual ambiguity, thus raising the need for a different representation for one of them. (MacEachren2012) published an extensive study on the graphical representations of uncertainty, such as glyph colors, textures, transparency, size and filling, these visual variables are presented in Figure 4.7 for completeness. According to their results, users accurately identify uncertainty when represented by the glyph's opacity and color saturation. Figures 4.8, 4.9, 4.10 and 4.11 show the results of experiments with the color saturation for property $W_p$, while figures 4.12, 4.13, 4.14 and 4.15 show the results of the same experiments, but for property $Q_o$. For both of these experiments we used fifty scenarios to avoid cluttering of the Time-lapsed LAMP chart.



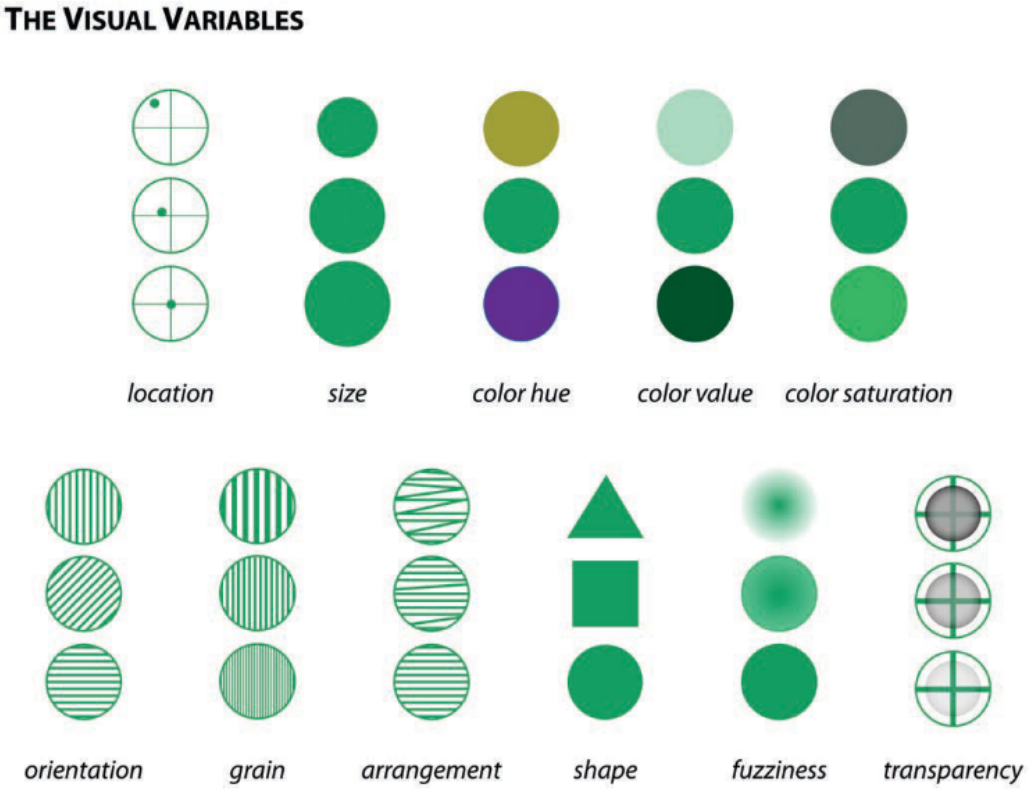
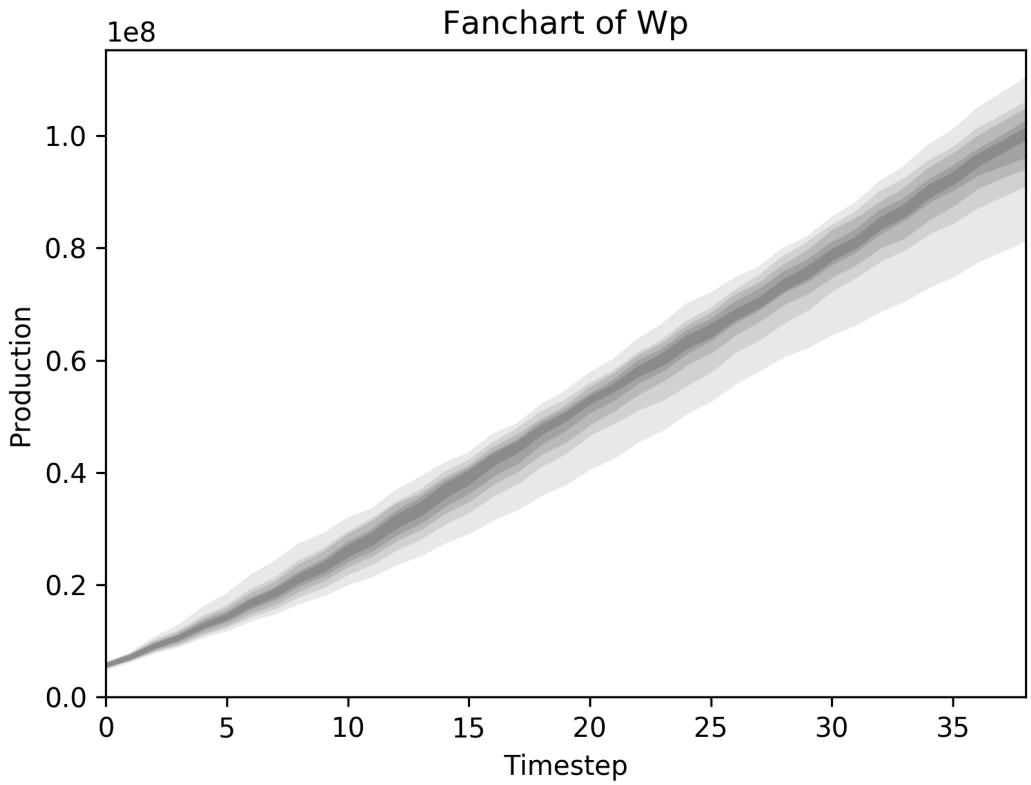
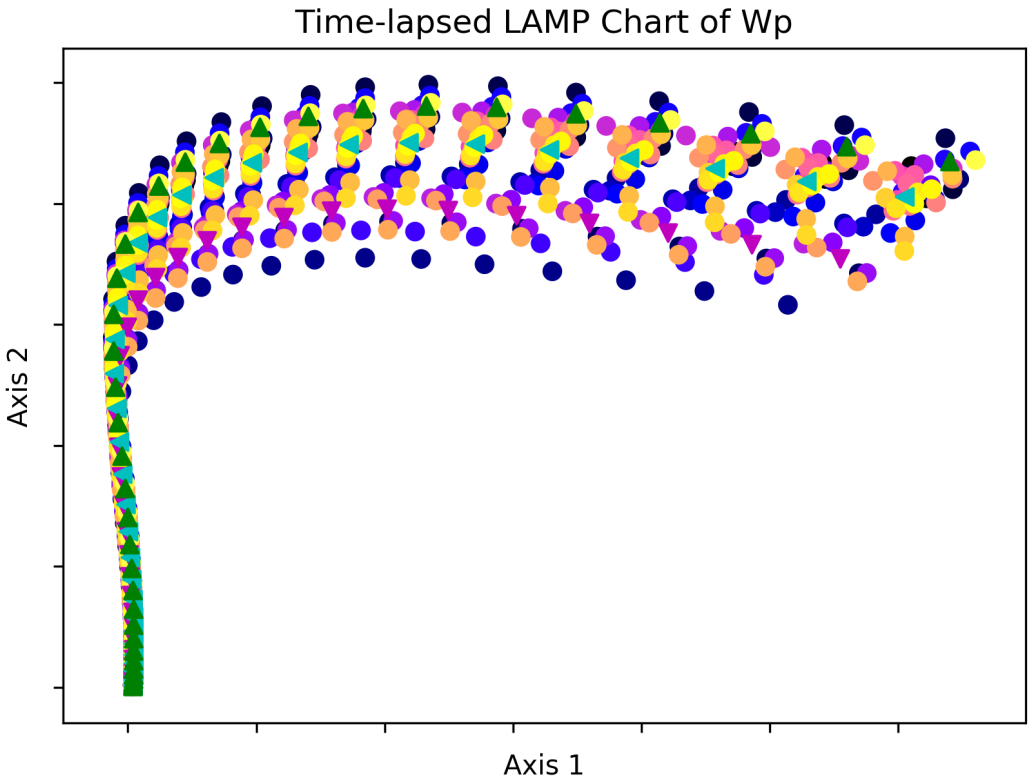



Figure 4.7: Visual variables categorized in the work of (MacEachren2012). Image taken from their work.

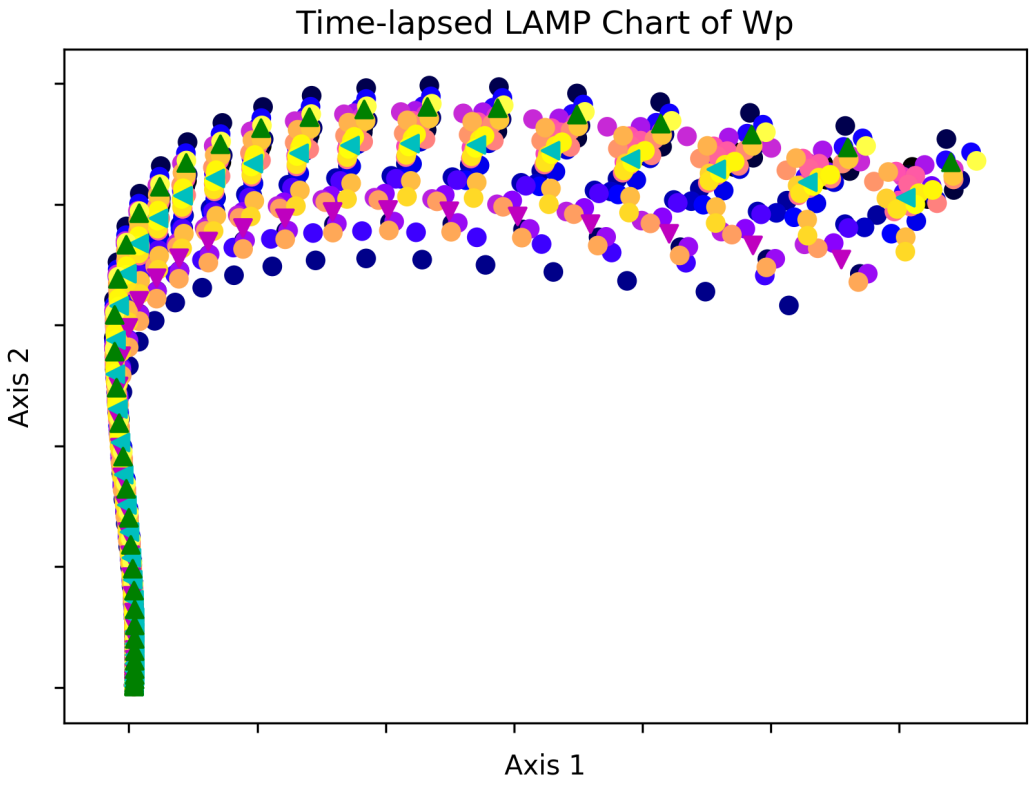



(a) Fanchart of $W_p$.

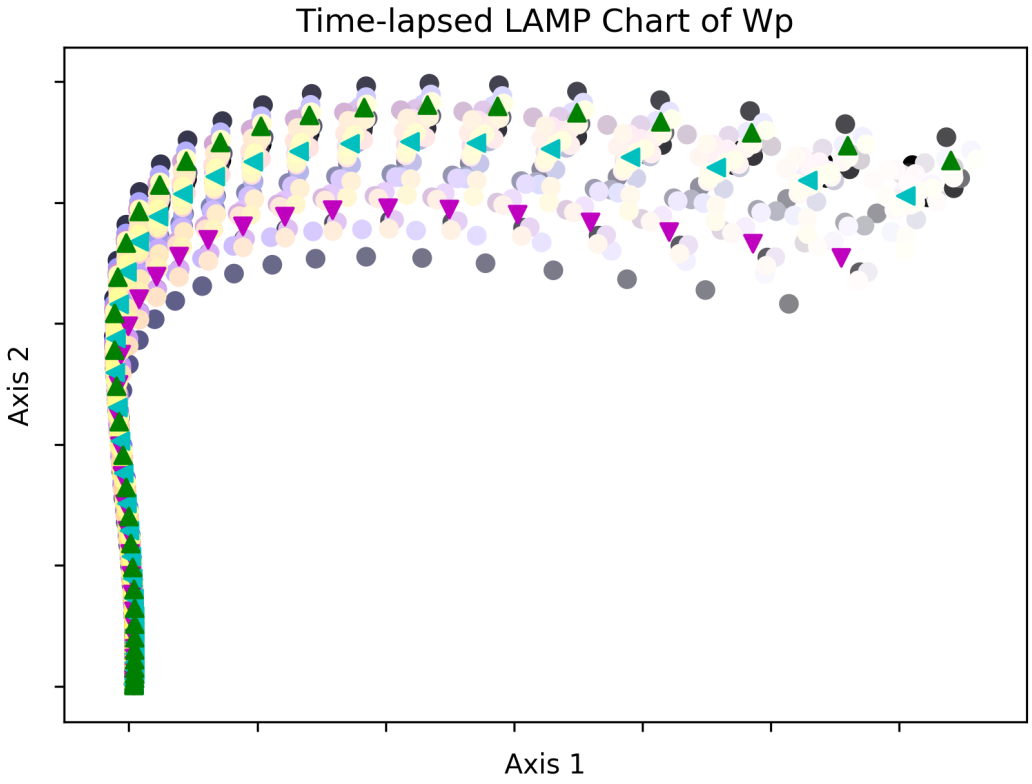



(b) Time-lapsed LAMP chart of $W_p$. Constant color saturation.

Figure 4.8: Fanchart (4.8a) and Time-lapsed LAMP chart of property $W_p$ using constant color saturation (4.8b).

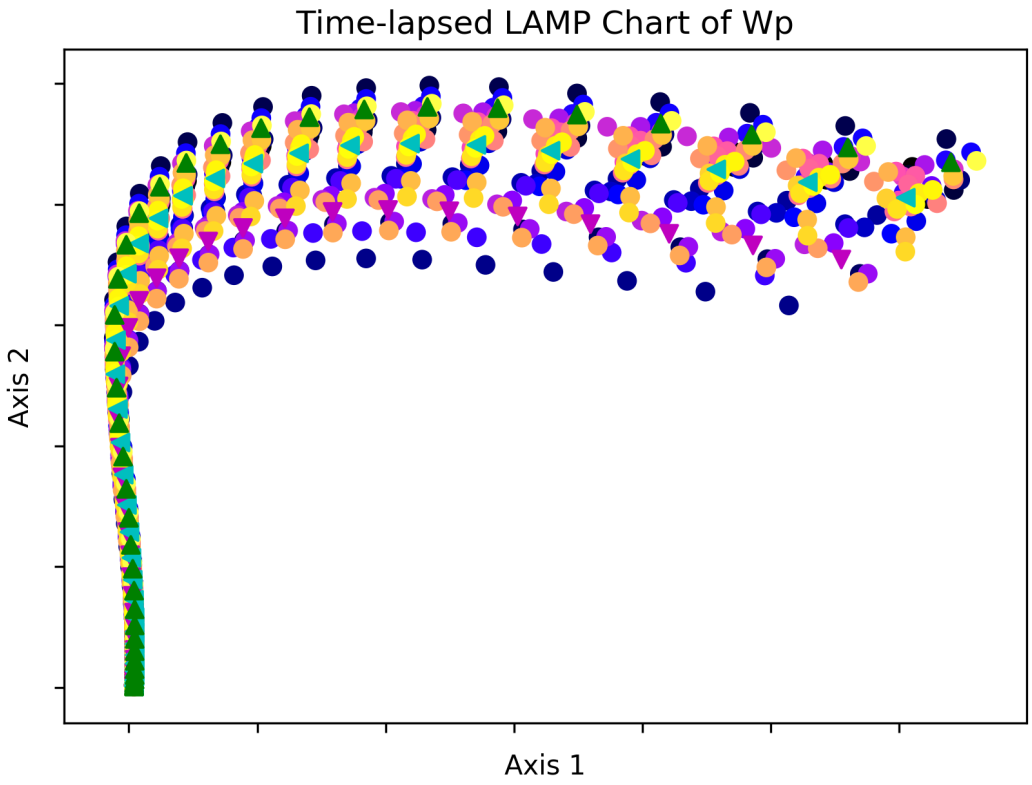



(a) Time-lapsed LAMP chart of $W_p$. Constant color saturation.

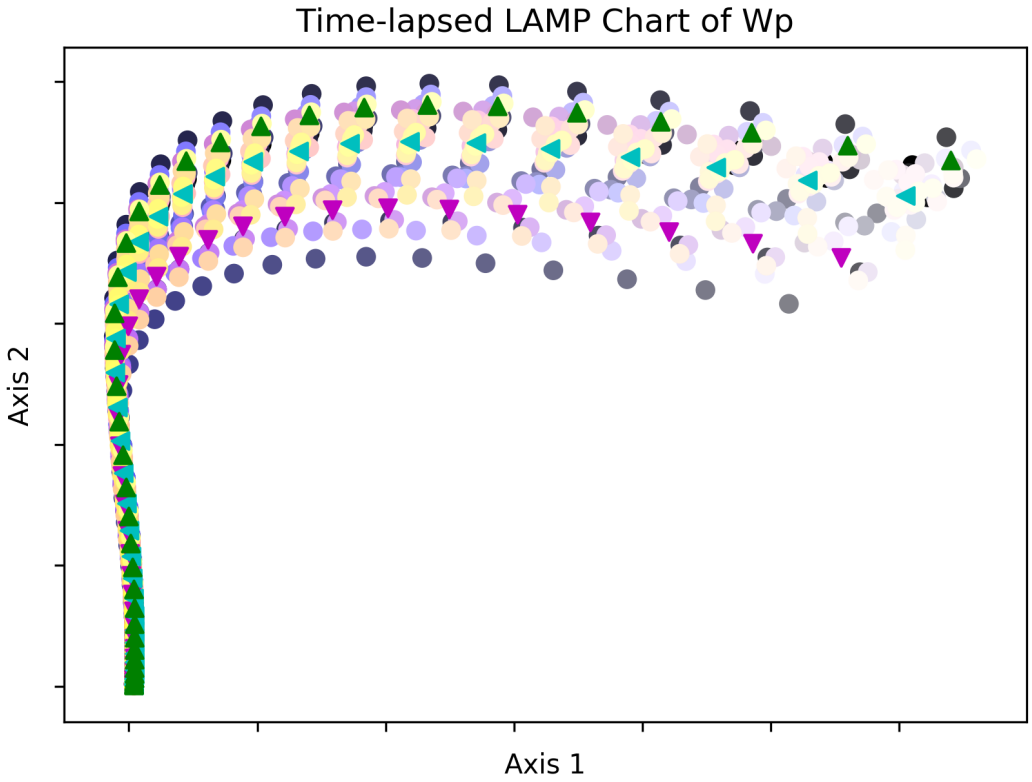



(b) Time-lapsed LAMP chart of $W_p$. Linearly decreasing color saturation.

Figure 4.9: Time-lapsed LAMP chart of property $W_p$. Comparison between constant color saturation (4.9a) and linearly decreasing color saturation (4.9b).

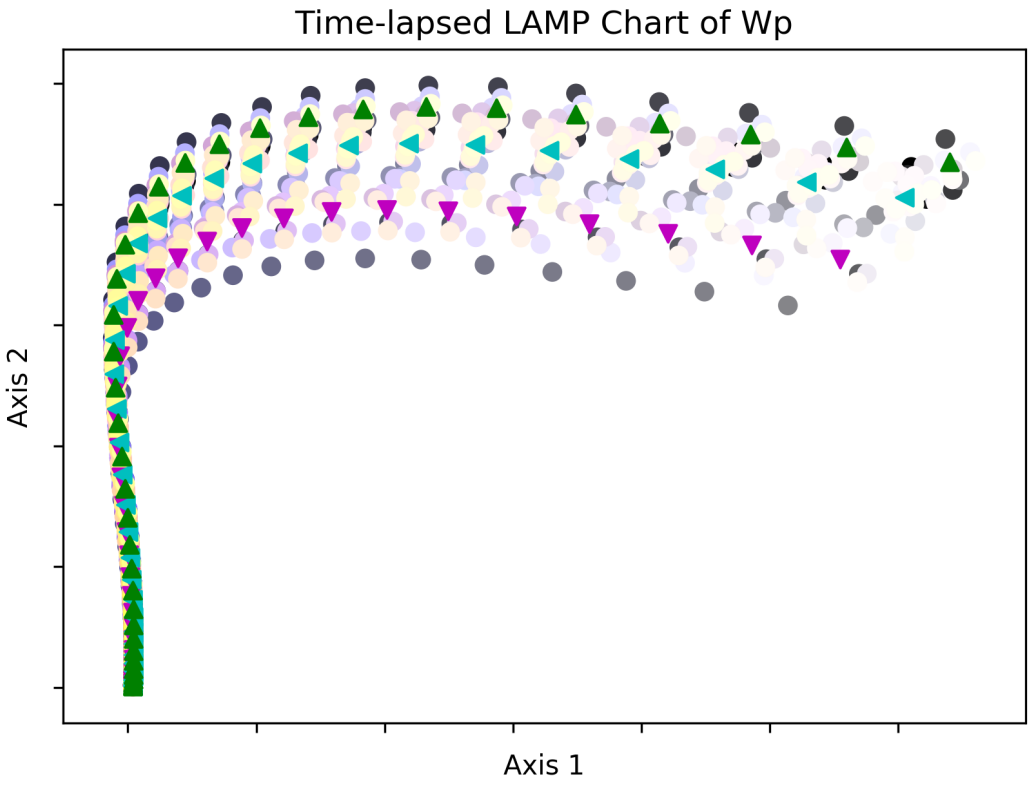



(a) Time-lapsed LAMP chart of $W_p$. Constant color saturation.

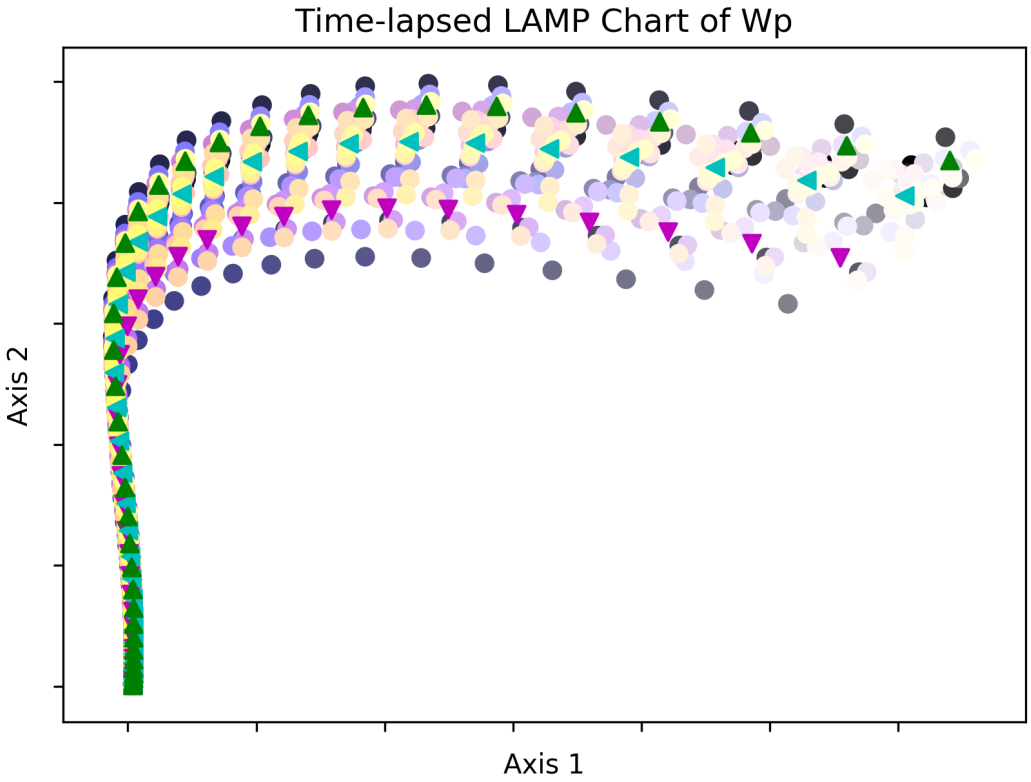



(b) Time-lapsed LAMP chart of $W_p$. Color saturation weighted by normalized data variance.

Figure 4.10: Time-lapsed LAMP chart of property $W_p$. Comparison between constant color saturation (4.10a) and color saturation weighted by normalized data variance (4.10b).

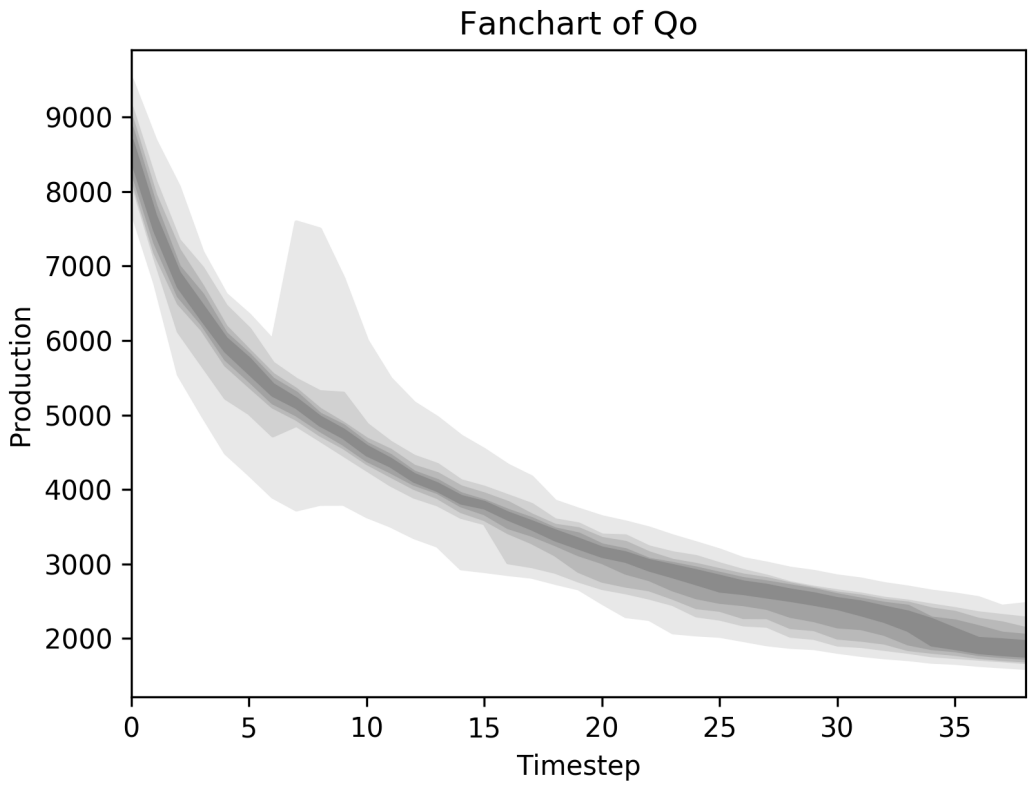



(a) Time-lapsed LAMP chart of $W_p$. Linearly decreasing color saturation.

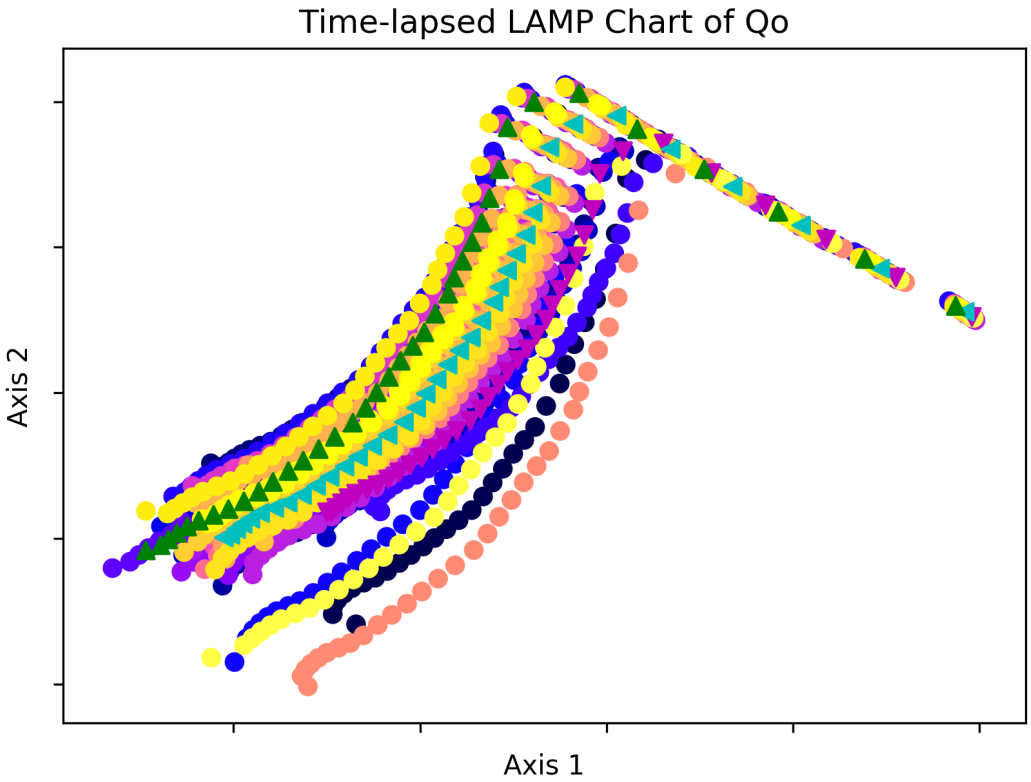



(b) Time-lapsed LAMP chart of $W_p$. Color saturation weighted by normalized data variance.

Figure 4.11: Time-lapsed LAMP chart of property $W_p$. Comparison between linearly decreasing color saturation (4.11a) and color saturation weighted by normalized data variance (4.11b).

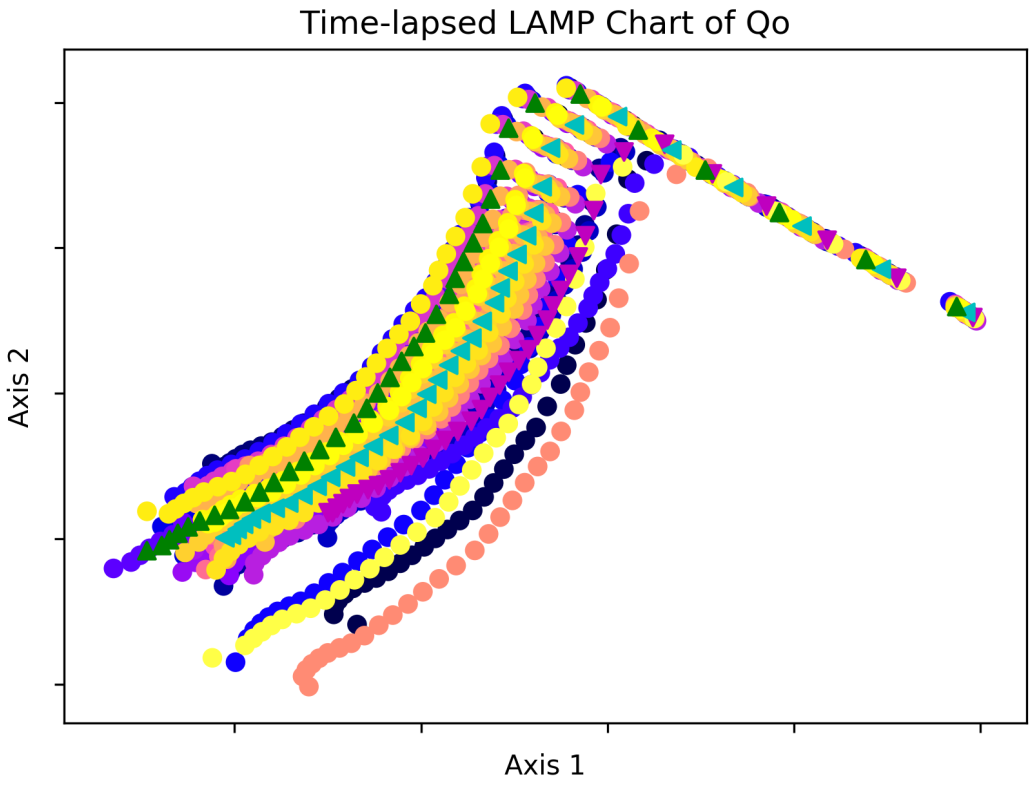



(a) Fanchart of $Q_o$.

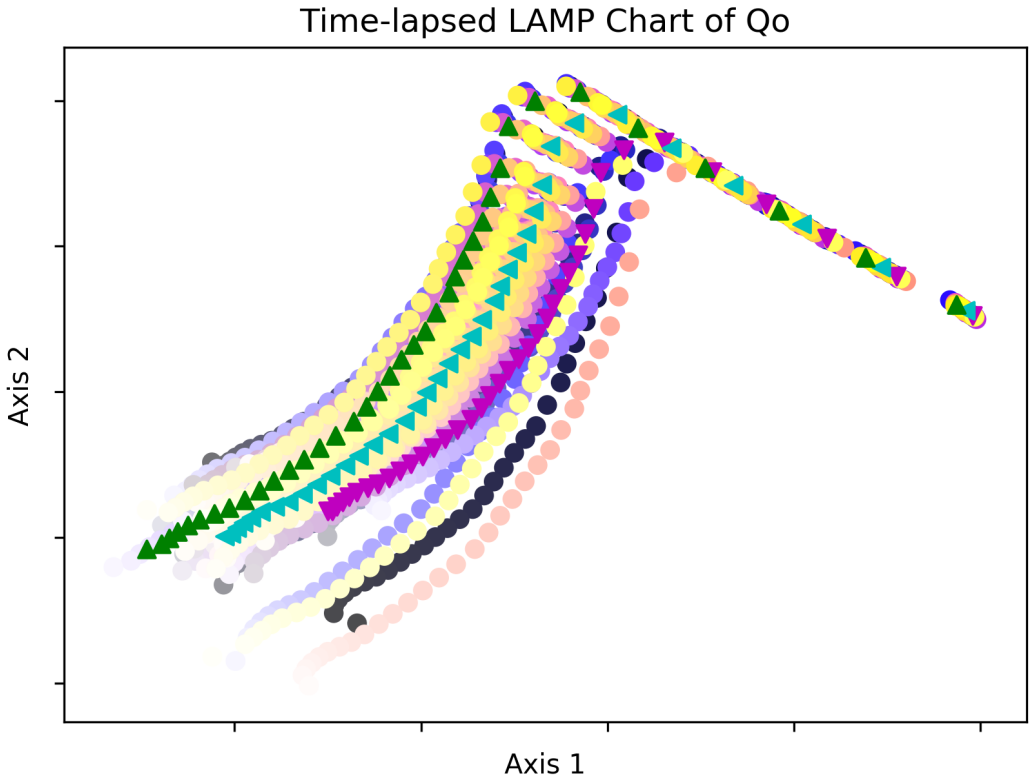



(b) Time-lapsed LAMP chart of $Q_o$. Constant color saturation.

Figure 4.12: Fanchart (4.12a) and Time-lapsed LAMP chart of property $Q_o$ using constant color saturation (4.12b).

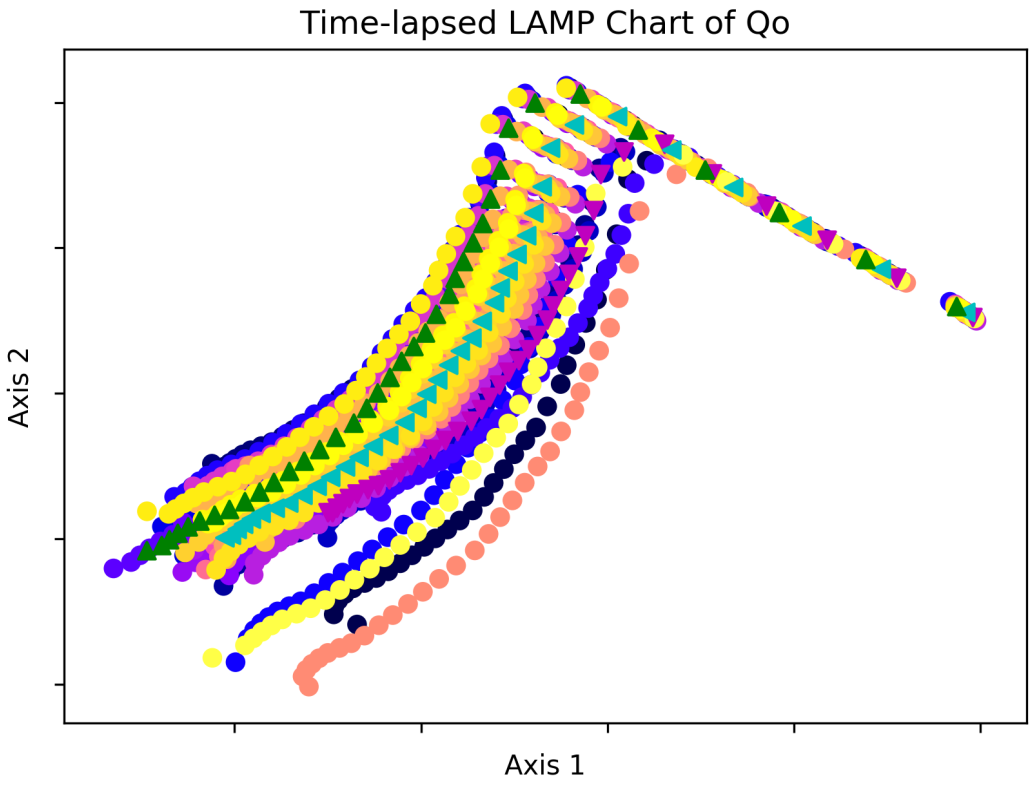



(a) Time-lapsed LAMP chart of $Q_o$. Constant color saturation.

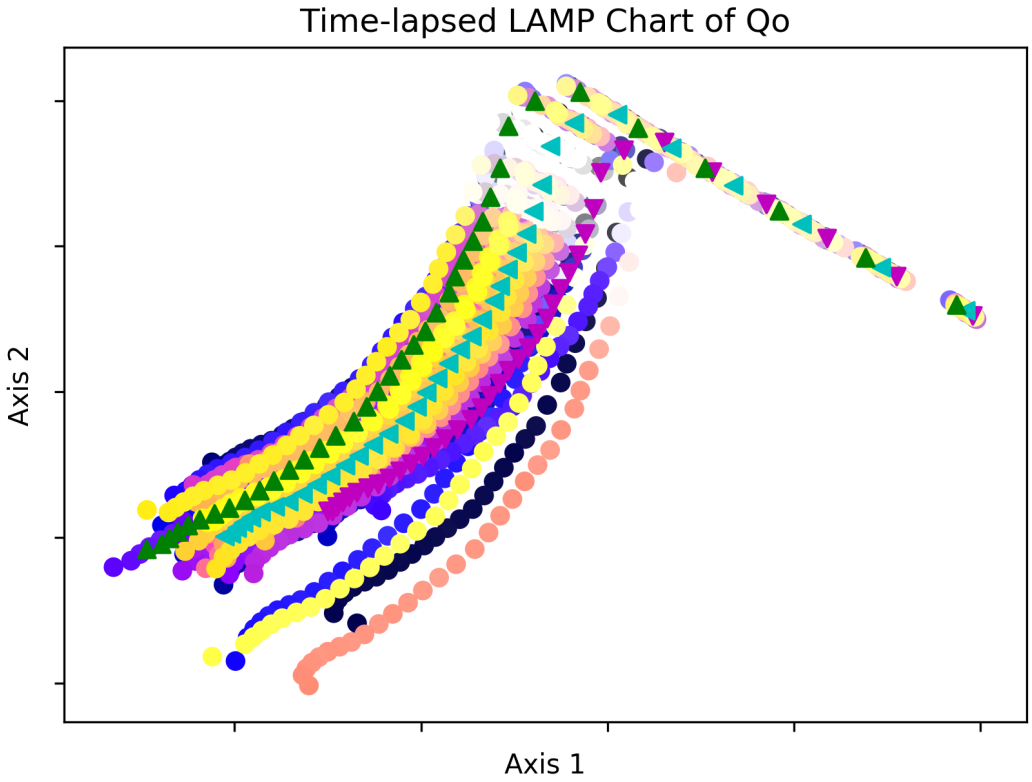



(b) Time-lapsed LAMP chart of $Q_o$. Linearly decreasing color saturation.

Figure 4.13: Time-lapsed LAMP chart of property $Q_o$. Comparison between constant color saturation (4.13a) and linearly decreasing color saturation (4.13b).

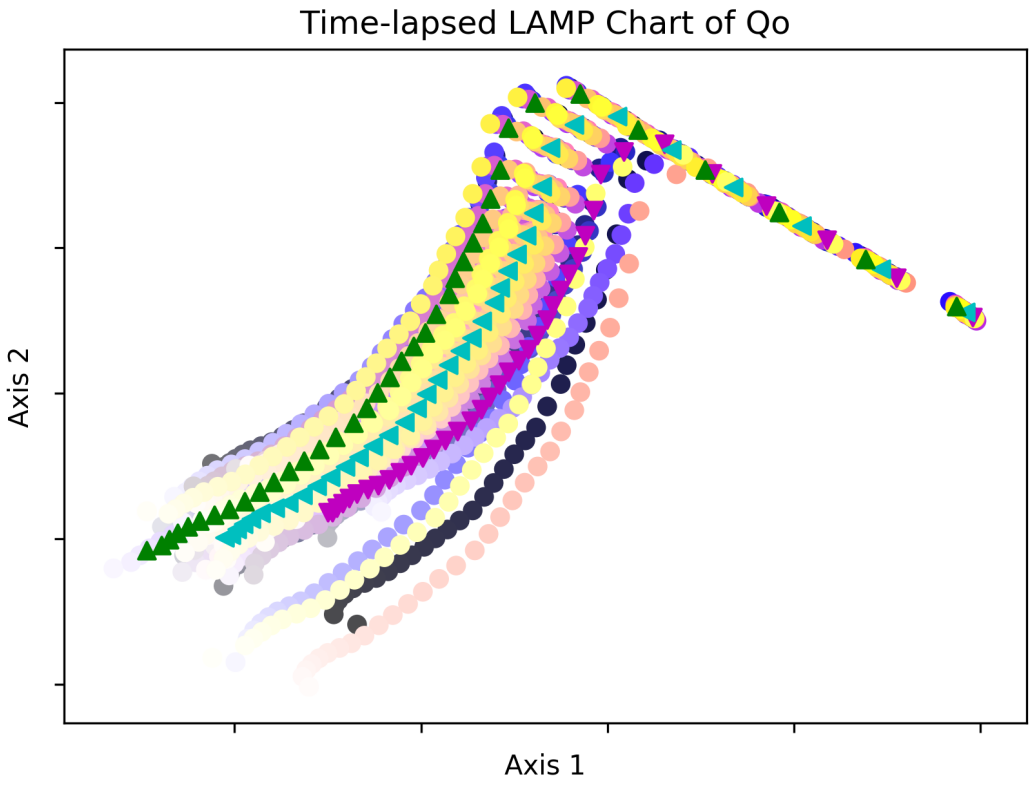



(a) Time-lapsed LAMP chart of $Q_o$. Constant color saturation.

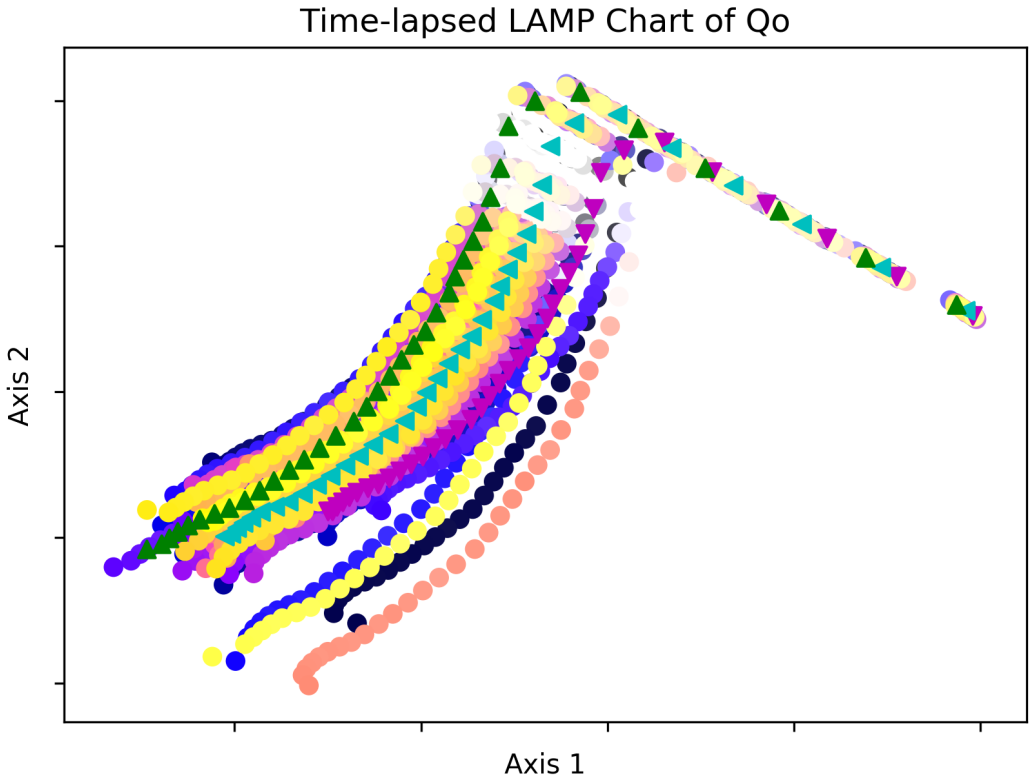



(b) Time-lapsed LAMP chart of $Q_o$. Color saturation weighted by normalized data variance.

Figure 4.14: Time-lapsed LAMP chart of property $Q_o$. Comparison between constant color saturation (4.14a) and color saturation weighted by normalized data variance (4.14b).




(a) Time-lapsed LAMP chart of $Q_o$. Linearly decreasing color saturation.




(b) Time-lapsed LAMP chart of $Q_o$. Color saturation weighted by normalized data variance.

Figure 4.15: Time-lapsed LAMP chart of property $Q_o$. Comparison between linearly decreasing color saturation (4.15a) and color saturation weighted by normalized data variance (4.15b).

According to (MacEachren2012) it seemed more sensible for users to encode the uncertainty as iconography, texture granularity, opacity and color saturation. However, iconography and texture granularity would add clutter to our chart, since the number of glyphs is fairly considerable even with a small number of scenarios. As for opacity, we use it to indicate a scenario highlighted by an user selection, leaving the color saturation as next best option. In this encoding, higher uncertainty is represented by less color saturation for the data at a certain time.

## 4.5 State-of-the-art Techniques

In Section 4.3 we presented results about our user-study, which indicated that our approach is a viable alternative for scenario reduction and data exploration in this context. In this section, we evaluate the performance of our proposed approach to other state-of-the-art and industry standard approaches.

We start by presenting details about the technique used in the oil & gas industry in Section 4.5.1, then, we present the approach proposed by (Shirangi2016) in Section 4.5.2. Finally, in Section 4.5.3 we compare the results of these two approaches with ours and discuss the results. For all tables in this section, bold-face rows indicate a scenario with smaller errors. All tests in this section use our full ensemble of two hundred scenarios as input.



### 4.5.1 Industry approach

The industry standard approach implemented in commercial software packages consists of selecting the scenarios with a value closest to the target percentile at a specified time for a single property. In order to compare Industry approach to ours, we mimicked it and tested with the cumulative oil and water production ($N_p$ and $W_P$ respectively). Table 4.1 presents the four scenarios selected by this approach for each percentile in question, as well as the sum of squared errors (SSE) and mean squared error (MSE) metrics. The scenarios with smallest error are shown in bold-face font. When a tie happens, the scenario with smallest euclidean distance is chosen. We can visually assess the proximity of the chosen scenarios using an MDS projection of the property data with the reference scenarios, as well as the scenarios chosen using each approach, which is shown in Figure 4.17.

Since this approach is widely used in the industry, we will use it as a reference for the results obtained by the other approaches described in our work.

Table 4.1: Scenarios and errors ($\times 10^{10}$) for the industry standard approach.

| Property | Percentile | Scenario | SSE | MSE |
|---|---|---|---|---|
| $N_p$ | $P_{10}$ | 172 | 1190.0 | 30.5 |
| | | 88 | 663.0 | 17.0 |
| | | 122 | 1660.0 | 42.6 |
| | | **36** | **312.0** | **7.9** |
| | $P_{50}$ | 131 | 162.0 | 4.1 |
| | | **4** | **162.0** | **4.1** |
| | | 26 | 786.0 | 20.1 |
| | | 90 | 236.0 | 6.0 |
| | $P_{90}$ | 96 | 975.0 | 25.0 |
| | | 100 | 690.0 | 17.7 |
| | | 127 | 925.0 | 23.7 |
| | | **132** | **526.0** | **13.5** |
| $W_p$ | $P_{10}$ | **23** | **7640.0** | **196.0** |
| | | 115 | 42000.0 | 1080.0 |
| | | 10 | 26900.0 | 690.0 |
| | | 19 | 27600.0 | 707.0 |
| | $P_{50}$ | 29 | 14200.0 | 364.0 |
| | | 153 | 6260.0 | 161.0 |
| | | **84** | **5840.0** | **150.0** |
| | | 88 | 11200.0 | 288.0 |
| | $P_{90}$ | 37 | 14300.0 | 367.0 |
| | | 57 | 22300.0 | 571.0 |
| | | 187 | 14000.0 | 358.0 |
| | | **28** | **2210.0** | **56.7** |

### 4.5.2 Clustering Approach

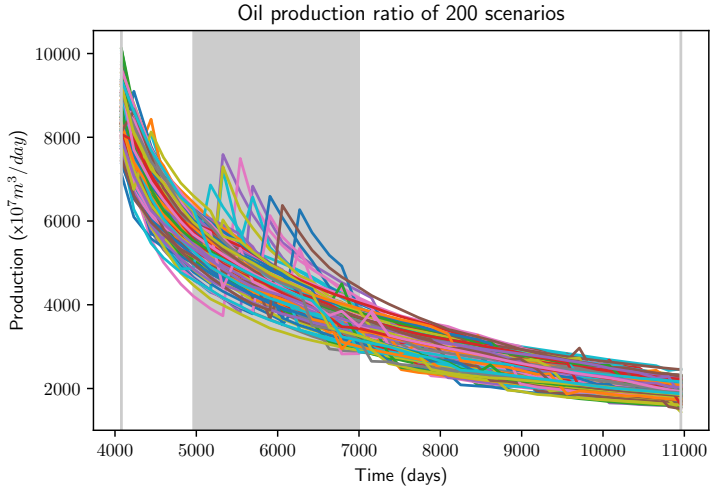
(Shirangi2016) proposes a general scenario reduction method based on clustering algorithms. Their approach consists clustering a low dimensional representation of an ensemble of scenarios, where each scenario is described by a feature vector encompassing both flow responses, such as oil and water productions, as well as geological parameters, such as field permeability and porosity.

They start by defining the low-dimensional representation for a scenario, called $\mathbf{r}$, which is composed by incremental production and injection data of each well of the scenario and described in Equation 4-1.

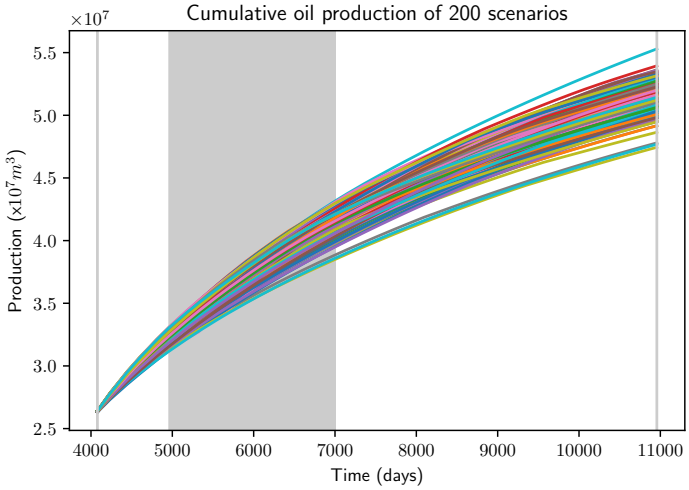
$$\mathbf{r}_i(\mathbf{x}, \mathbf{m}_i) = \left[\mathbf{q}_i^1 \mathbf{q}_i^2 \dots \mathbf{q}_i^{N_w}\right]^T \tag{4-1}$$

where $\mathbf{x}$ is a well-parameter vector, such as locations and bottom-hole pressure values. $\mathbf{q}_i^k$ is the $k$-th well production/injection data for the $i$-th scenario, and $N_w$ is the number of wells in the model.

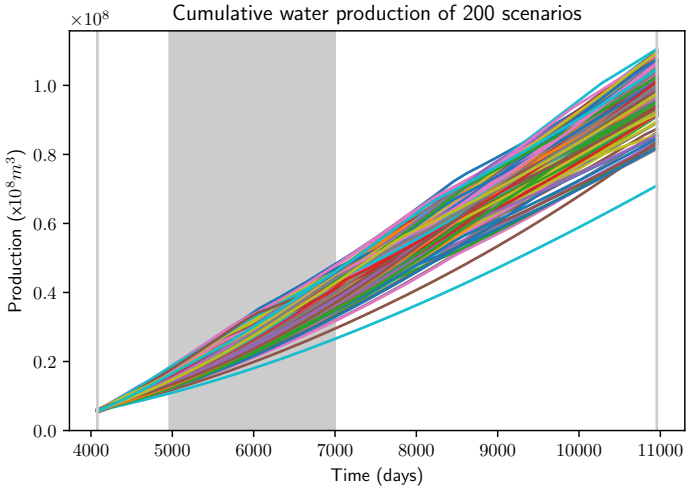
In order to compose the $\mathbf{q}_i^k$ vectors, a number of time intervals must be chosen. How to choose such intervals depends on the data itself. It must be noted, however, that the number of intervals must be kept a relatively small in order to generate an actual lower-dimensionality representation of a scenario. For our tests, we chose three intervals based on the oil production ratio ($Q_o$) of our scenarios. Figure 4.16a shows the $Q_o$ of 200 scenarios of our ensemble. The gray vertical span between simulation days 4900 and 7000 constitutes an interesting range, since some scenarios have a spike in oil production between those times before returning to stable values. Figures 4.16b and 4.16c show the cumulative oil and water productions for our scenarios with the same time range highlighted.




(a)




(b)




(c)

Figure 4.16: Oil production ratio, cumulative oil and water production of 200 scenarios in our ensemble. The highlighted time range was based on the oil production ratio (4.16a) due to interesting production spikes in that range.

Based on the features presented in Figure 4.16a, we selected the simulation days 4079 (first forecast day, marked as time step 0), 4962, 7001, which composes the beginning and end of the highlighted time range, and finally, day 10957, which is the last simulation day.

Having defined the time steps, we can build a matrix $\mathbf{Z}_f = [\mathbf{r}_1\mathbf{r}_2 \dots \mathbf{r}_{N_R}]^T$, where $N_R$ is the total number of scenarios in the ensemble, and then perform the next steps in (Shirangi2016) approach, detailed below:

1. Specify the number of representative scenarios $n_r$ to be selected of $N_R$;
2. Build the feature matrix $\mathbf{Z}_f$;
3. Apply the k-means clustering algorithm with $n_r$ clusters;
4. For each cluster found by k-means:
   - Search for the scenario closest to the centroid found by k-means;
5. Return these scenarios as the representatives.

Our implementation of (Shirangi2016) differs in two main aspects from the original approach: (i) we do not use individual well data in our flow feature matrix $\mathbf{Z}_f$; (ii) we do not change the well parameters $\mathbf{x}$. Regarding the first point, the well data is aggregated, thus leading to one feature array for each scenario. As for the second point, our goal is not well-parameter estimation, thus, we do not alter those parameters. We also do not include the geological property feature matrix $\mathbf{Z}_p$ in our tests; this is equivalent of setting the $\alpha$ weighting parameter to 0 in their approach, thus, assigning full weight to the flow-based features (Shirangi2016).

For the k-means clustering, we set the percentiles scenarios as initial centroids. We also set the number of maximum iterations to 10000 and the tolerance to $1e-6$. The resulting scenarios, plus errors and distance from the closest percentile, are presented in Table 4.2. Figure 4.17 shows the MDS projection of the properties and the selected scenarios. For the cumulative oil production, the resulting scenarios are actually closer to the $P_{10}$ and $P_{50}$ scenarios, while for the cumulative water production, the selected scenarios are more spread across the projection plane, but, with the exception of the $P_{50}$, they are not closer to the references than the scenarios chosen by the other approaches.

Table 4.2: Scenarios and errors ($\times 10^{10}$) for (Shirangi2016) of each property under consideration.

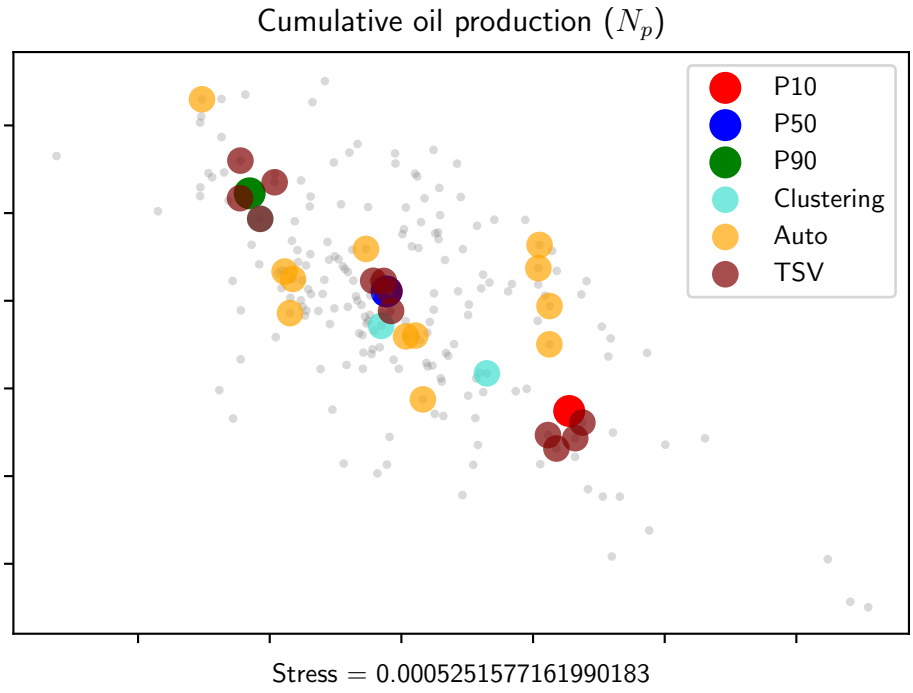
| **Property** | **Scenario** | **SSE** | **MSE** |
|---|---|---|---|
| | 39 | 5540.0 | 142.0 |
| $N_p$ | 104 | 1790.0 | 46.0 |
| | 159 | 96.6 | 2.4 |
| | 39 | 9320.0 | 239.0 |
| $W_p$ | 104 | 1270.0 | 32.5 |
| | 159 | 52000.0 | 1330.0 |

### 4.5.3 Comparison and Discussions

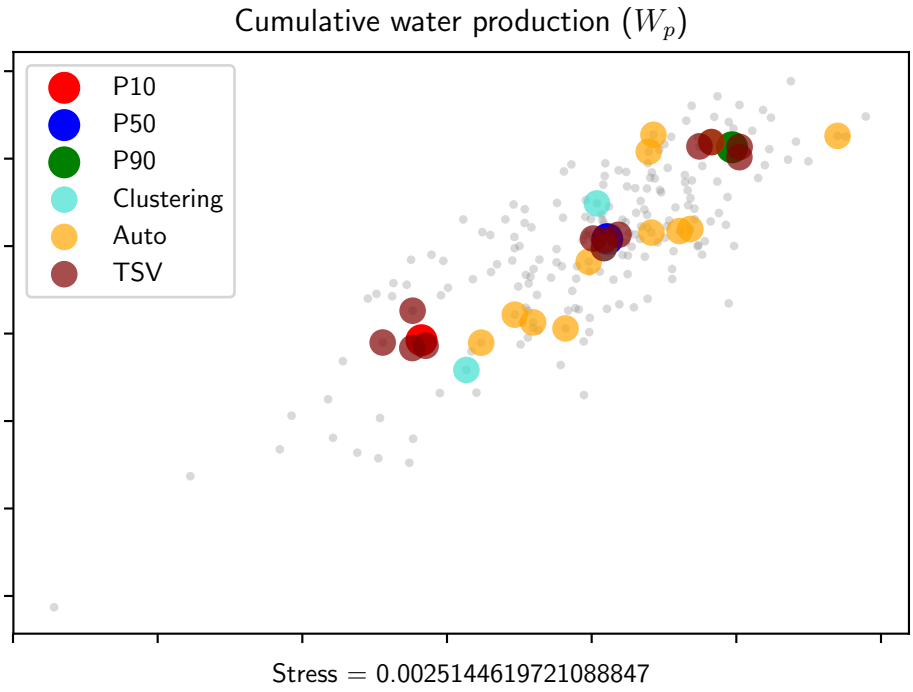
Table 4.3 presents the error measures for the scenarios obtained by using our approach for selecting scenarios. For each percentile, we manually selected the four visually closest scenarios. Figure 4.17 shows an MDS projection of $N_p$ and $W_p$ data with the scenarios selected by each approach marked in different colors, as well as the reference percentiles. The stress of the projections is low, in the order of $10^{-5}$ and $10^{-3}$, which indicates that the distances between the scenarios are well represented in this low-dimensional representation.

Table 4.3: Scenarios and errors ($\times 10^{10}$) for scenarios selected using our approach of each property.

| **Property** | **Percentile** | **Scenario** | **SSE** | **MSE** |
|---|---|---|---|---|
| $N_p$ | $P_{10}$ | **186** | **78.2** | **2.0** |
| | | 73 | 108.0 | 2.7 |
| | | 147 | 130.0 | 3.3 |
| | | 58 | 160.0 | 4.1 |
| | $P_{50}$ | **194** | **23.4** | **0.6** |
| | | 112 | 25.3 | 0.6 |
| | | 197 | 37.1 | 0.9 |
| | | 45 | 39.3 | 1.0 |
| | $P_{90}$ | **66** | **25.1** | **0.6** |
| | | 159 | 96.6 | 2.4 |
| | | 54 | 105.0 | 2.6 |
| | | 156 | 113.0 | 2.9 |
| $W_p$ | $P_{10}$ | **129** | **826.0** | **21.2** |
| | | 149 | 1730.0 | 44.3 |
| | | 117 | 3750.0 | 96.2 |
| | | 152 | 5630.0 | 144.0 |
| | $P_{50}$ | **46** | **680.0** | **17.4** |
| | | 104 | 1270.0 | 32.5 |
| | | 30 | 1420.0 | 36.3 |
| | | 61 | 1510.0 | 38.7 |
| | $P_{90}$ | 109 | 2080.0 | 53.2 |
| | | **16** | **1590.0** | **40.7** |
| | | 28 | 2210.0 | 56.7 |
| | | 43 | 3560.0 | 91.2 |




(a) MDS projection of property $N_p$ for two hundred scenarios in our ensemble.




(b) MDS projection of property $W_p$ for two hundred scenarios in our ensemble.

Figure 4.17: MDS projections of the cumulative oil (4.17a) and water productions (4.17b) of our ensemble. The reference percentiles are colored in red ($P_{10}$), blue ($P_{50}$) and green ($P_{90}$). The scenarios selected by each approach are colored as follows: (Shirangi2016) in turquoise, Industry standard in orange and our approach in maroons.

In order to compare the error measures of all approaches, we present their summarized results in Table 4.4. The scenarios with smallest error metrics are presented in bold-face.

Table 4.4: Summary of the error measures for all approaches presented (errors $\times 10^{10}$).

| Property | Percentile | Scenario | SSE | MSE | Approach |
|---|---|---|---|---|---|
| $N_p$ | $P_{10}$ | 36 | 312.0 | 7.9 | Auto |
| | | 39 | 554.0 | 142.0 | Shirangi |
| | | **186** | **78.0** | **2.0** | **TSV** |
| | $P_{50}$ | 4 | 162.0 | 4.1 | Auto |
| | | 104 | 1790.0 | 46.0 | Shirangi |
| | | **194** | **23.4** | **0.6** | **TSV** |
| | $P_{90}$ | 132 | 526.0 | 13.5 | Auto |
| | | 159 | 98.8 | 2.4 | Shirangi |
| | | **66** | **25.1** | **0.6** | **TSV** |
| $W_p$ | $P_{10}$ | 23 | 7640.0 | 196.0 | Auto |
| | | 39 | 9320.0 | 239.0 | Shirangi |
| | | **129** | **826.0** | **21.2** | **TSV** |
| | $P_{50}$ | 84 | 5840.0 | 150.0 | Auto |
| | | 104 | 1270.0 | 32.3 | Shirangi |
| | | **46** | **680.0** | **17.4** | **TSV** |
| | $P_{90}$ | 28 | 2210.0 | 56.7 | Auto |
| | | 159 | 52000.0 | 1330.0 | Shirangi |
| | | **16** | **1590.0** | **40.7** | **TSV** |

The error metrics in Table 4.4 show that our approach results in scenarios with consistently lower error values compared to the industry standard and (Shirangi2016)'s approach. However, (Shirangi2016)'s approach obtains representative scenarios across several properties, while our approach currently handles a single property at a time. Their approach may also use simulation parameters in order to guide the process, while ours do not provide support for that.

Figures 4.18, 4.19 and 4.20 show a visual comparison of scenarios selected by each approach for the cumulative oil production. For this property, our approach yields visually better scenarios for all three percentiles. For $P_{10}$, show in Figure 4.18, the industry approach results in an overall good scenario as well (`UNISIM-I-H_036`), as shown in the Fanchart, Bump and Distance charts. Figure 4.19 shows the results using $P_{50}$ as a baseline. In this case the industry approach also yields a good scenario (`UNISIM-I-H_004`), while the clustering approach results in a scenario visually far from the baseline. Finally, for the

baseline $P_{90}$ in Figure 4.20 the three approachs result in excelent scenarios. The clustering approach results in a scenario visually closer to the baseline when compared to the industry approach. Our approach yields a scenario closer in than both, but by a small margin, as shown in the Distance Chart.

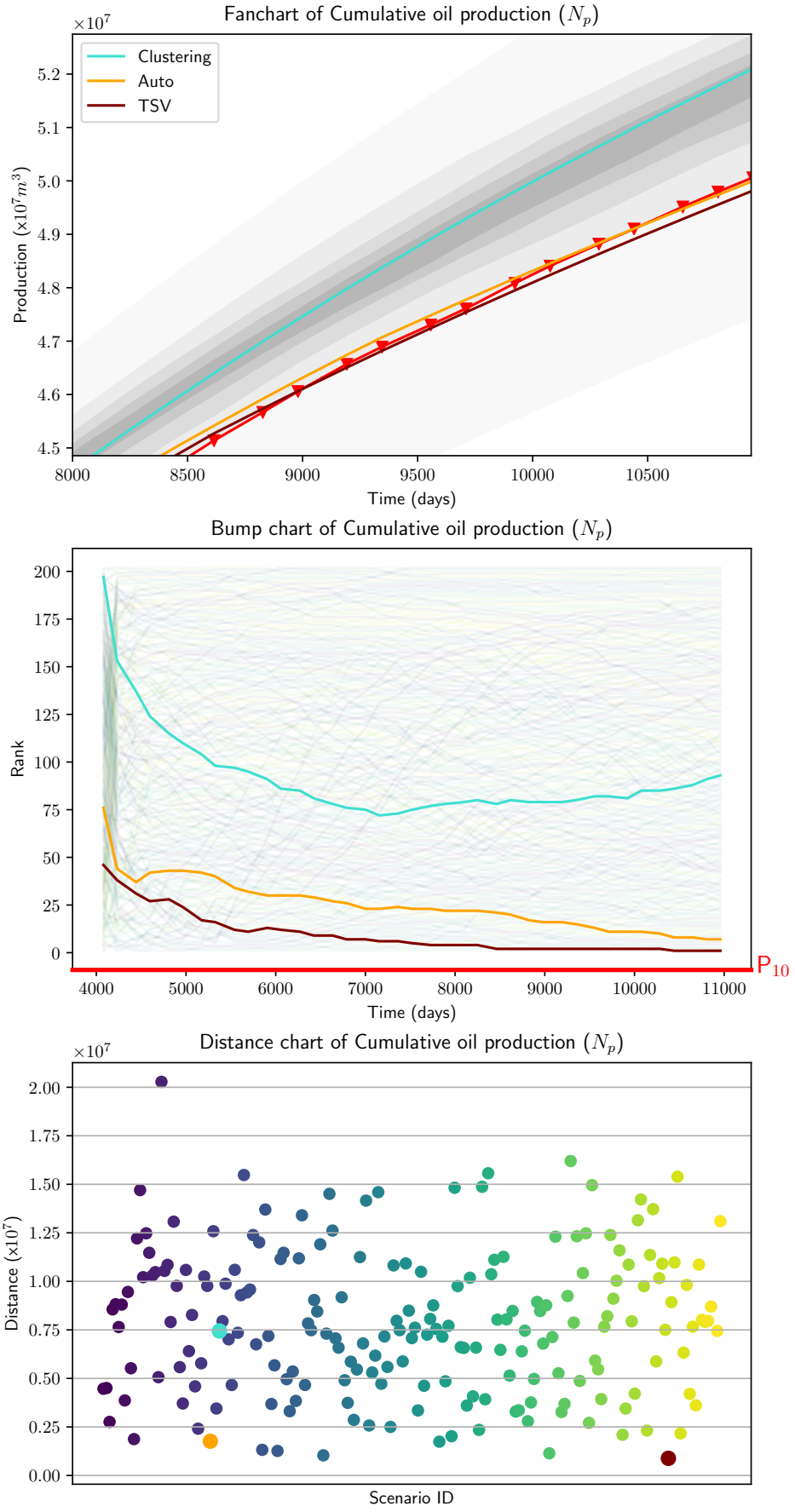



Figure 4.18: Resulting scenarios for the $P_{10}$ baseline of the $N_p$ property.

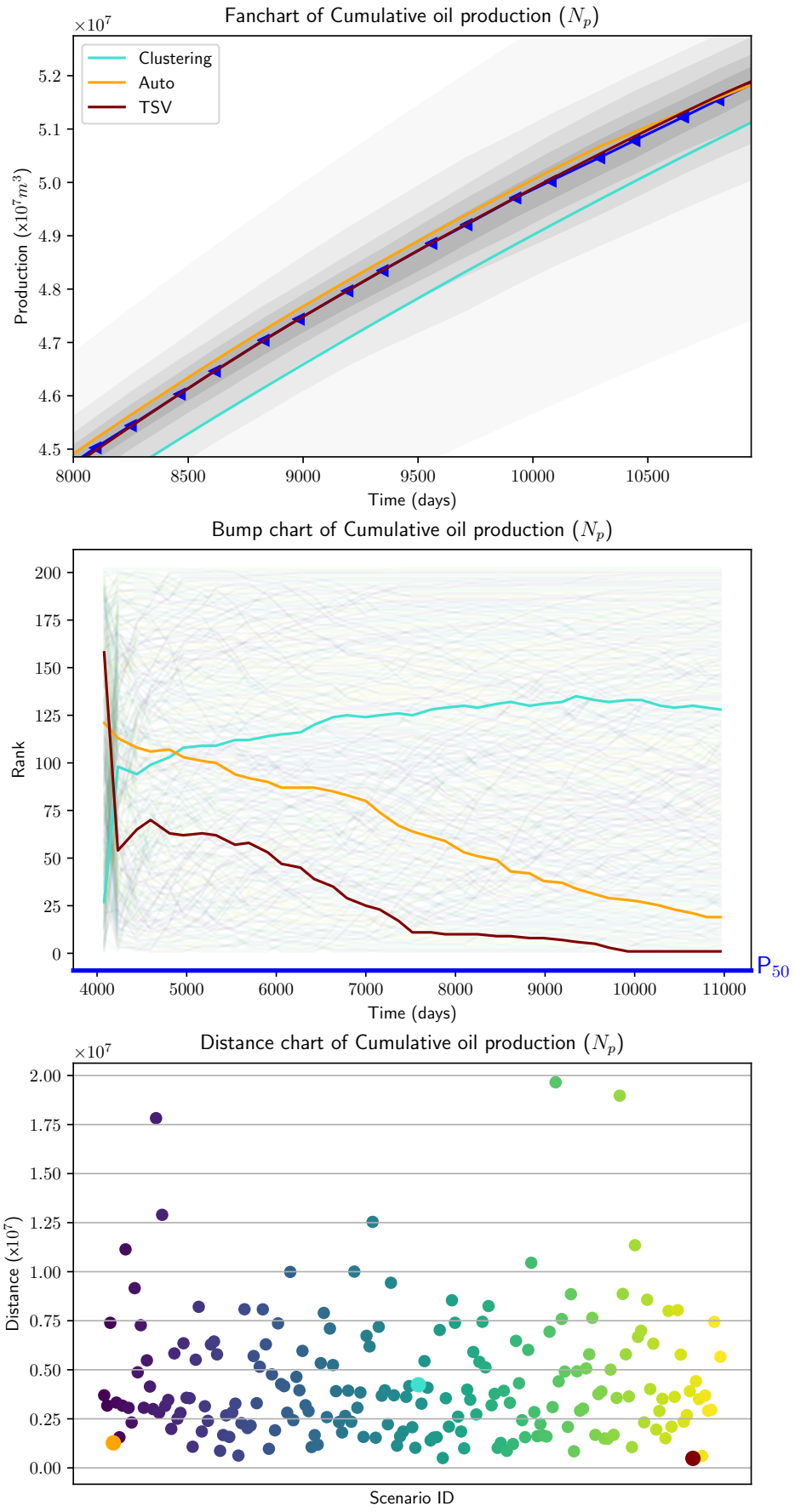



Figure 4.19: Resulting scenarios for the $P_{50}$ baseline of the $N_p$ property.

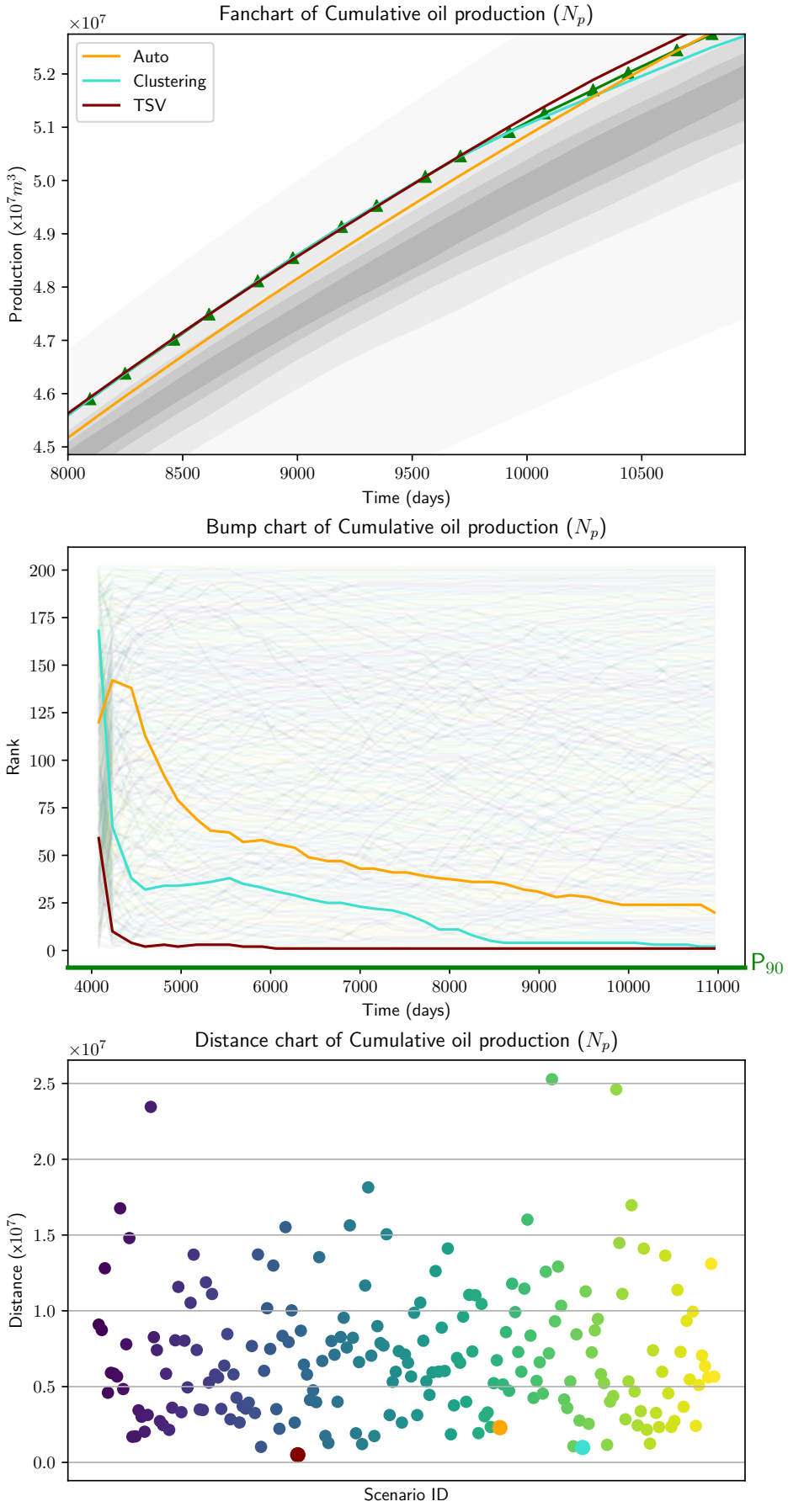



Figure 4.20: Resulting scenarios for the $P_{90}$ baseline of the $N_p$ property.

The same visual comparison done for the cumulative oil production, was performed for the cumulative water production, shown in figures 4.21, 4.22 and 4.23. For this property, the three approaches resulted in competitive scenarios, especially for the $P_{10}$ and $P_{50}$ baselines. The visual results of the $P_{50}$ baseline, in Figure 4.22, if the simulation had stopped at approximately 7500 days, an user of our approach may have selected a scenario visually less adherent to the reference when compared to the other approaches, however, as the simulation advances beyond that time, the scenario chosen with our approach (`UNISIM-I-H_046`) gets closer to the reference, and maintains this behavior until the simulation's end. In this particular case, the difference in error measures can be explained by the magnitude of the data towards the end of the simulation. Table 4.5 shows the range of values at some simulation days for the property $W_p$.

Table 4.5: Minimum, maximum and range of production values for selected timesteps of $W_p$ (values $\times 10^5$.

| **Simulation days** | **Min. production** | **Max. production** | **Difference** |
|---:|---:|---:|---:|
| 4079 | 263 | 264 | 1 |
| 8096 | 415 | 471 | 56 |
| 10957 | 474 | 553 | 79 |

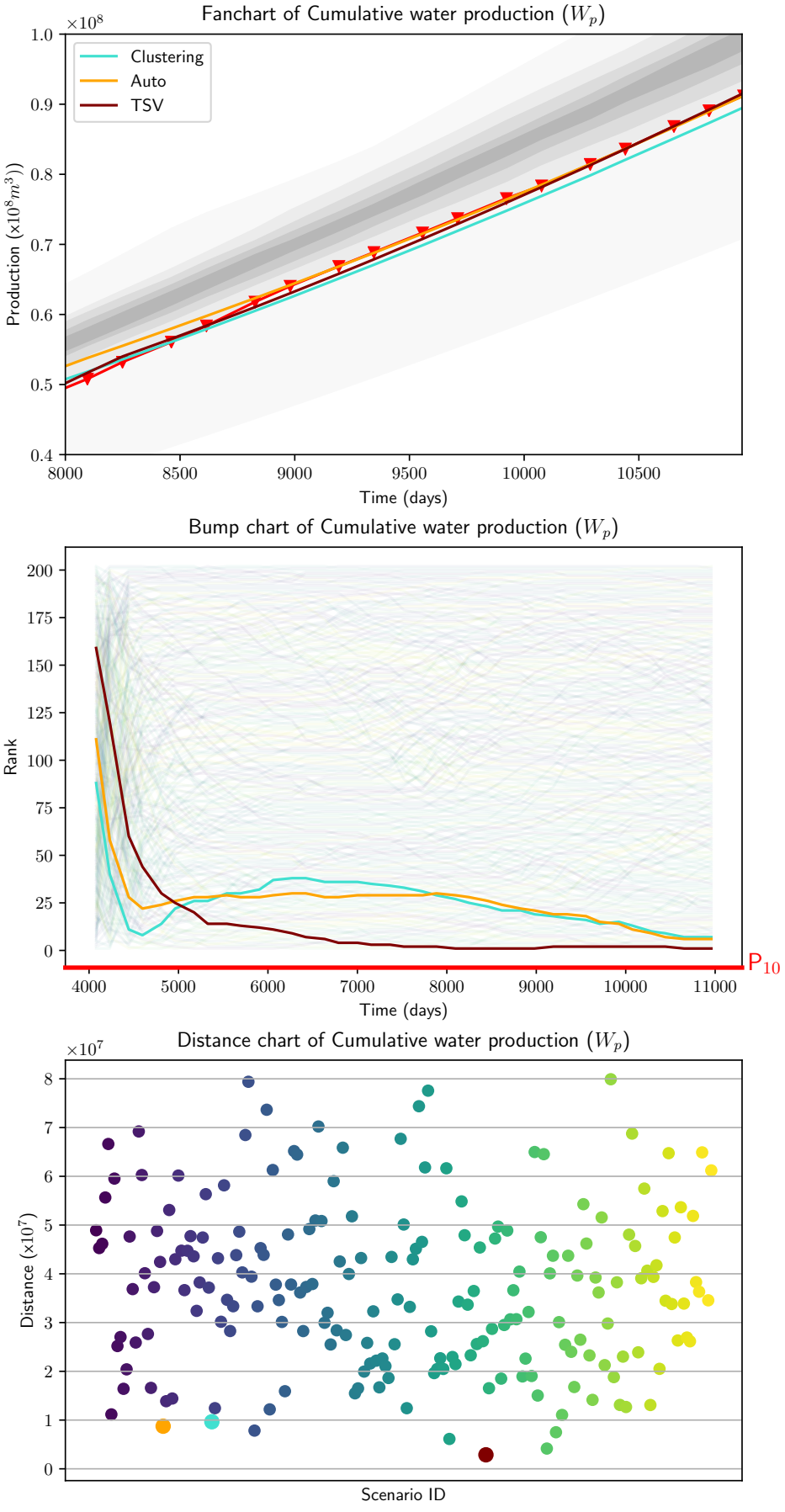



Figure 4.21: Resulting scenarios for the $P_{10}$ baseline of the $W_p$ property.

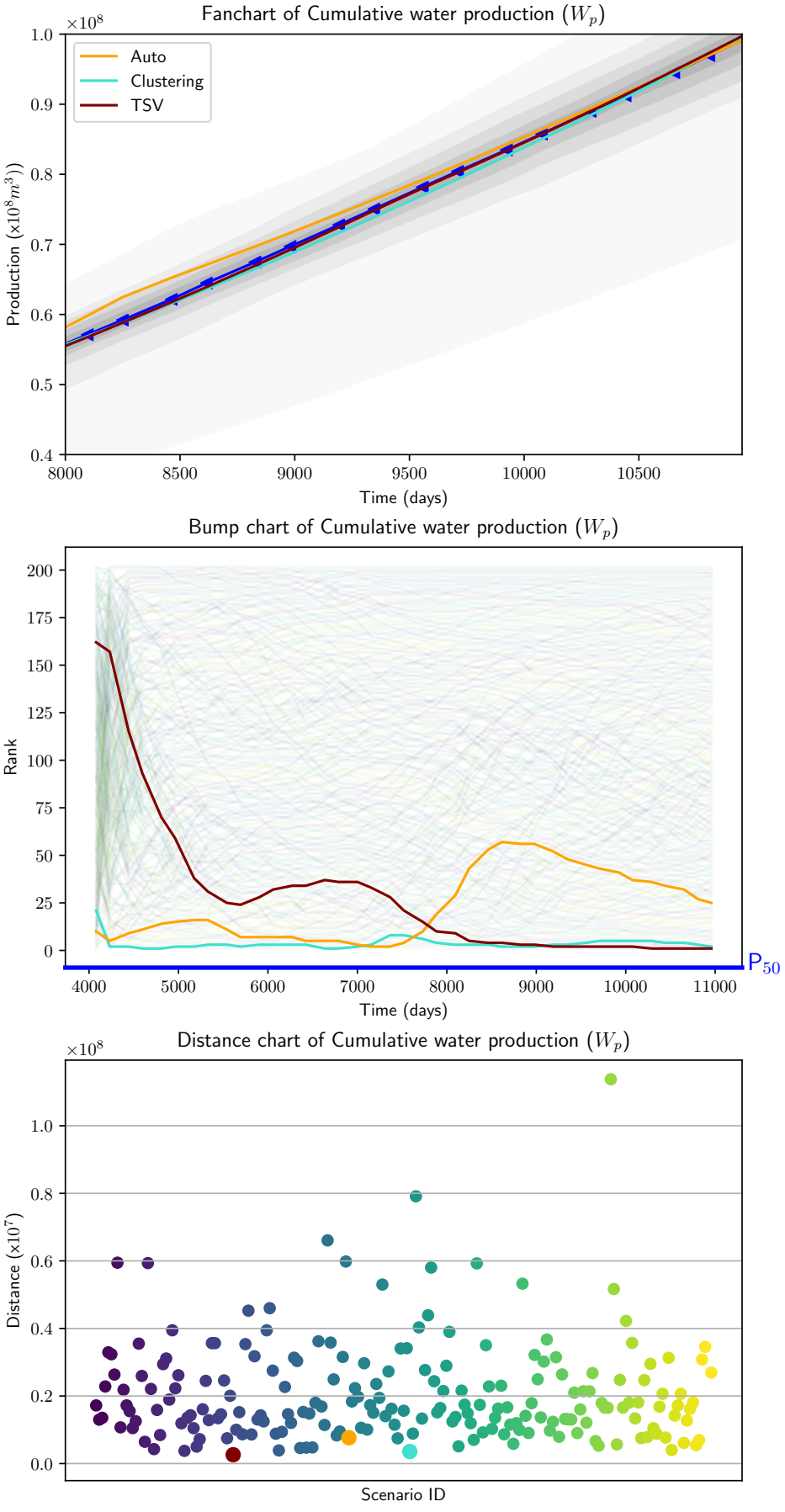



Figure 4.22: Resulting scenarios for the $P_{50}$ baseline of the $W_p$ property.

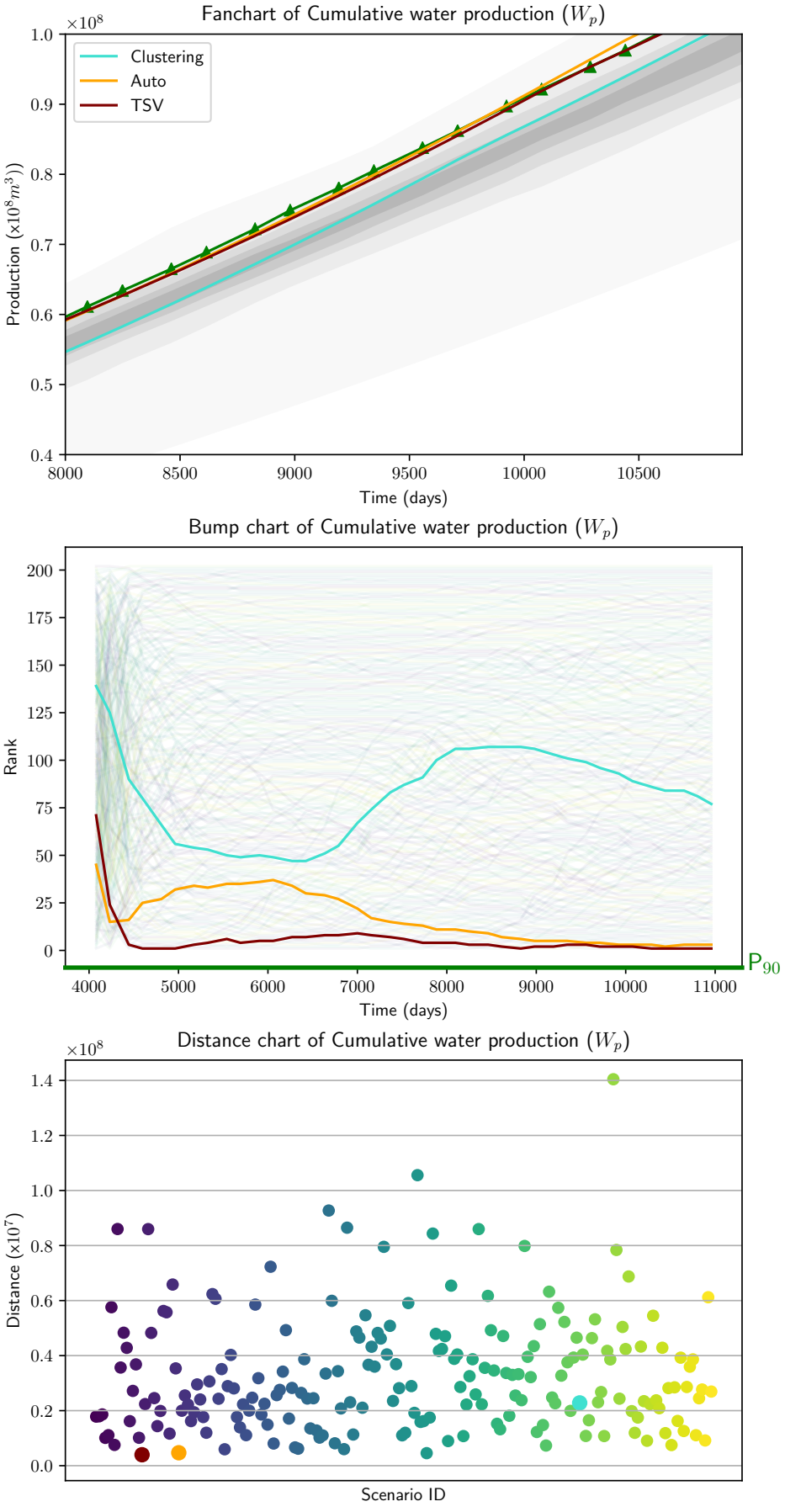



Figure 4.23: Resulting scenarios for the $P_{90}$ baseline of the $W_p$ property.

In order to perform a proper scenario reduction, the user must consider the goal of the simulation process itself, i.e., the problem being solved and what analyses will be done afterwards. (Shirangi2016) introduced their approach for selecting scenarios considering various well configurations across a range of geophysical properties (Shirangi2016). In such contexts, their approach yields good results. Our approach, on the other hand, shows good results in the context of temporal scenarios, when the objective function is the proximity between a set of scenarios and a reference. The industry standard approach chooses scenarios with a value close to a reference goal at the last simulation time, disregarding the intermediate values, making it a good approach for selecting scenarios that will reach a certain goal independent on the scenarios evolution. In addition, the visual adherence results for $W_p$, shown in figures 4.21, 4.22 and 4.23 and the errors shown in Table 4.4 indicate that more than one approach should be used for a given problem and the results compared, in order to yield better subsets of scenarios for further analysis.

# 5
# Conclusions

This chapter presents the concluding remarks about our work. We list the publications originated from this work in Section 5.1, discuss our contributions in Section 5.2, and present directions for future works in Section 5.3.

In this thesis, we presented a novel graphical approach to scenario reduction on time series ensembles. We evaluated the feasibility of our proposal by performing an empirical study with a series of potential users, both experienced in the area and not. By observing their interactions and interviewing them afterwards, we obtained valuable insights on the usefulness of our proposal. Following from the results of our previous publication (Schardong2018) we have expanded our work in two aspects: (i) use glyph sizes to represent the time in the Time-lapsed LAMP chart; (ii) in the same chart, encode the uncertainty inherent to the data. The first expansion was done in order to fix an issue related to the abstraction created by employing multidimensional-projections using time-varying data. The second expansion aims to help users to quickly identify time ranges with high variance in the data.

Besides the user study, we also compared our results with other approaches in the literature and industry. (Shirangi2016) proposes to select a set of representative scenarios under different well configurations. Their approach handles several simulation parameters and responses simultaneously, thus, selecting representative scenarios based on multiple criteria. The Inducstry approach selects scenarios with cumulative production closest to the target references at the end of the simulation. Both approaches are well suited depending on the post-processing tasks and the data itself. In our tests, our approach selected scenarios with consistently smaller error when compared to both the Industry and Clustering approaches. It must be stated, however, that this does not mean that our approach is better, only that it yields good scenarios when the objective function is the proximity to a reference in a context where the scenarios' evolution must be taken into consideration.

## 5.1
## Publications

Our work resulted in the following publication:

SCHARDONG, G. G.; RODRIGUES, A. M.; BARBOSA, S. D. ; LOPES, H.. Visual interactive support for selecting scenarios from time-series ensembles. Decision Support Systems, 113:99–107, sep 2018.

## 5.2 Contributions

Our work advanced the state-of-the-art by proposing a graphical approach to explore and perform scenario reduction tasks in a decision support context. Our approach handles time series data, which is an important contribution, especially in the context of oil & gas industry, where the most widely used approaches only handle a single point in time.

We also proposed adaptations on two visualization techniques: (i) the Local Affine Multidimensional Projection (Joia2011); and (ii) the Bump chart (Tufte1990). In the former, we proposed a way to trace the evolution of the ensemble by comparing the distances between all elements at different simulation times. In the latter, we proposed a transformation of the input temporal data to a cumulative, ordinal form in order to visualize it as a series of rankings that evolve with time. We did so by calculating the distances between each time series and a reference series and ordering it, thus, composing the a series' rank at each simulation time.

Besides the research contributions, our work aimed to solve a problem faced by PETROBRAS during the decision making process for oil reservoir management. Our goals were defined by problems raised during the project entitled "*Visualização e quantificação de incertezas de um conjunto de simulações de reservatório*", with National Petroleum Agency (ANP) number 18008-3, namely, select percentile scenarios given an ensemble of production scenarios. A prototype of our solution was implemented and delivered as conclusion for project "*Simulação estocástica, modelagem, otimização e análise estatística de dados de poços aplicados à avaliação de formações*", with ANP number 17987-9, since the first project had ended in the year 2016, before a final version of our prototype was fully functional.

## 5.3 Future Works

An interesting research direction would be to use metrics from the risk management area, such as Value at Risk (VaR) and Continuous Value at Risk (CVaR) in order to calculate how representative are the chosen scenarios. These metrics may also be incorporated into a framework for assessment of scenario reduction approaches, which in turn may lead to the development of Ensemble

Scenario Reduction approaches, essentially combining the results of several algorithms in order to propose the best possible subset of scenarios.

Another interesting work would be to incorporate the evolution of the scenarios in optimization approaches such as the one proposed by (Sarma2013). They propose a minimax algorithm to search for scenarios with minimal distance of the percentiles, while maximizing their spread in the parameter uncertainty space. This approach may be used in place of the clustering step in (Shirangi2016), thus creating an approach to find representative scenarios in various well configurations, while maximizing the spread in the parameter space.

There is a body of literature in the most diverse areas that investigate the problem of finding a representative subset of elements in a larger set, such as: computational geometry (Agarwal2005), stochastic and robust optimization areas (Arpon2018). Drawing from the computational geometry area, the notion of representative scenarios is closely related to the notion of Core-sets (Agarwal2005), which is defined as a subset $Q \in S$ so that solving the underlying problem in $Q$ yields an approximate solution for $S$. In this regard, a natural extension of our work would be to model the scenario reduction problem as a coreset problem and adapt the techniques developed by the computational geometry community to the scenario reduction problem.

In the robust optimization area one of the goals is to find effective scenarios, i.e., scenarios whose removal from the set causes the objective function value to fall from its optimal value (Rahimian2018). In this context, the goal is to find such scenarios while removing the ineffective scenarios from the set. The set of effective scenarios are representative of the larger set, and can be used to study the underlying uncertainties of the problem being solved. All these approaches may be used to solve scenario reduction problems in different contexts, and the decision of which approach to use may not be straightforward. In this regard, our approach can be adapted into a visual analytics and decision support framework, where the results of several scenario reduction approaches can be visually and objectively compared. This in turn may fill a gap we noticed in the literature during the course of our work, which is the lack of visual exploration approaches for scenario reduction, thus leaving room for future works in this direction.

Other areas for further research involve better representation of uncertainty and time in the visualizations employed. (MacEachren2012) compared several visual representations of uncertainty and concluded that opacity and saturation convey the uncertainty information more clearly than glyph size and texture. We did explore some representations, such as opacity and satu-

ration; however, other aspects may be explored. (Wallner2000) studied error propagation in affine combinations of complex bodies from a geometric perspective. The time and uncertainty representations in the Time-lapsed LAMP chart may be improved using the concepts introduced in their work, such as mapping the uncertainty at time step $t$ as convex regions of size proportional to $var(t)$ and interpolating these regions using B-splines. This way, the uncertainty at each time may be represented not only by the glyph color saturation, but also by size and deformation of the convex regions and interpolated polygons. Another chart that may benefit from these concepts is the Distance chart. A simple but effective way to incorporate time in this chart would be to incrementally present the distance to the objective function for each scenario. The closer a time step is to the last time step, the less transparent it is, until the last time step, which is presented with full opacity. This way, an user may quickly assess the contribution of each time step to the overall distance to the objective function. The variance of distances may also be presented as a boxplot overlaid in the Distance chart.

With a prototype of our work delivered to PETROBRAS as product of a project, we can measure the quality of decisions made using their percentile selection methods compared to our proposal, refine our prototype, and eventually, launch it as a full featured reservoir management decision-making tool.

# Bibliography


[Agarwal2005] AGARWAL, P. K.; HAR-PELED, S. ; VARADARAJAN, K. R.. **Geometric Approximation via Coresets**. Combinatorial and Computational Geometry MSRI Publications, 2005.

[Aigner2007] AIGNER, W.; BERTONE, A.; MIKSCH, S.; TOMINSKI, C. ; SCHUMANN, H.. **Towards a conceptual framework for visual analytics of time and time-oriented data**. In: SIMULATION CONFERENCE, 2007 WINTER, p. 721–729, 2007.

[Alencar2012] ALENCAR, A. B.; BÖRNER, K.; PAULOVICH, F. V. ; DE OLIVEIRA, M. C. F.. **Time-aware visualization of document collections**. In: PROCEEDINGS OF THE 27TH ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING - SAC '12, SAC '12, p. 997–1004, New York, New York, USA, 2012. ACM Press.

[Alrefaei2007] ALREFAEI, M. H.; ALMOMANI, M.. **Subset selection of best simulated systems**. Journal of the Franklin Institute, 344(5):495–506, 2007.

[Andrienko2011] ANDRIENKO, G.; ANDRIENKO, N.; KEIM, D.; MACEACHREN, A. M. ; WROBEL, S.. **Editorial: Challenging problems of geospatial visual analytics**. J. Vis. Lang. Comput., 22(4):251–256, Aug. 2011.

[Armstrong2013] ARMSTRONG, M.; NDIAYE, A.; RAZANATSIMBA, R. ; GALLI, A.. **Scenario Reduction Applied to Geostatistical Simulations**. Mathematical Geosciences, 45(2):165–182, feb 2013.

[Armstrong2014] ARMSTRONG, M.; VINCENT, A.; GALLI, A. ; MEHEUT, C.. **Genetic algorithms and scenario reduction**. Journal of the Southern African Institute of Mining and Metallurgy, 114(3):237–244, 2014.

[Arpon2018] ARPÓN, S.; HOMEM-DE MELLO, T. ; PAGNONCELLI, B.. **Scenario reduction for stochastic programs with conditional value-at-risk**. Mathematical Programming, 170(1):327–356, Jul 2018.

[Avansi2015] AVANSI, G. D.; SCHIOZER, D. J.. **UNISIM-I: Synthetic Model for Reservoir Development and Management Applications**. International Journal of Modeling and Simulation for the Petroleum Industry, 9(1), 2015.

[Becker1987] BECKER, R. A.; CLEVELAND, W. S. ; HILL, M.. **Brushing Scatterplots**. Technometrics, 29(2):127–142, 1987.

[Britton1998] BRITTON, E.; FISHER, P. ; WHITLEY, J.. **The Inflation Report projections: understanding the fan chart**. Bank of England Quarterly Bulletin, (Feb):30–37, 1998.

[Buja1991] BUJA, A.; MCDONALD, J.; MICHALAK, J. ; STUETZLE, W.. **Interactive data visualization using focusing and linking**. In: PROCEEDING VISUALIZATION '91, p. 156–163,. IEEE Comput. Soc. Press, 1991.

[Cheng2016] CHENG, X.; COOK, D. ; HOFMANN, H.. **Enabling Interactivity on Displays of Multivariate Time Series and Longitudinal Data**. Journal of Computational and Graphical Statistics, 25(4):1057–1076, 2016.

[Demir2014] DEMIR, I.; DICK, C. ; WESTERMANN, R.. **Multi-charts for comparative 3D ensemble visualization**. IEEE Transactions on Visualization and Computer Graphics, 20(12):2694–2703, 2014.

[DiDomenica2007] DOMENICA, N. D.; MITRA, G.; VALENTE, P. ; BIRBILIS, G.. **Stochastic programming and scenario generation within a simulation framework: An information systems perspective**. Decision Support Systems, 42(4):2197 – 2218, 2007. Decision Support Systems in Emerging Economies.

[Dupacova2003] DUPAČOVÁ, J.; GRÖWE-KUSKA, N. ; RÖMISCH, W.. **Scenario reduction in stochastic programming**. Mathematical Programming, 95(3):493–511, 2003.

[Eler2009] ELER, D. M.; PAULOVICH, F. V.; DE OLIVEIRA, M. C. F. ; MINGHIM, R.. **Topic-based coordination for visual analysis of evolving document collections**. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON INFORMATION VISUALISATION, p. 149–155, 2009.

[Emerick2013] EMERICK, A. A.; REYNOLDS, A. C.. **Ensemble smoother with multiple data assimilation**. Computers & Geosciences, 55:3 – 15, 2013.

[Ester1996] ESTER, M.; KRIEGEL, H. P.; SANDER, J. ; XU, X.. **A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**. In: PROCEEDINGS OF THE 2ND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, p. 226–231, 1996.

[Faloutsos1995] FALOUTSOS, C.; LIN, K.-I.. **FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets**. In: PROCEEDINGS OF THE 1995 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, volumen 24, p. 163–174, 1995.

[Gower2004] GOWER, J. C.; DIJKSTERHUIS, G. B.. **Procrustes problems**. Número 30. Oxford University Press on Demand, 2004.

[Growe-Kuska2003] GRÖWE-KUSKA, N.; HEITSCH, H. ; RÖMISCH, W.. **Scenario reduction and scenario tree construction for power management problems**. In: 2003 IEEE BOLOGNA POWERTECH, volumen 3, p. 152–158, 2003.

[Heitsch2003] HEITSCH, H.; RÖMISCH, W.. **Scenario reduction algorithms in stochastic programming**. Computational Optimization and Applications, 24(2-3):187–206, 2003.

[Heitsch2009] HEITSCH, H.; RÖMISCH, W.. **Scenario tree modeling for multistage stochastic programs**. Mathematical Programming, 118(2):371–406, may 2009.

[Hummel2013] HUMMEL, M.; OBERMAIER, H.; GARTH, C. ; JOY, K. I.. **Comparative visual analysis of lagrangian transport in cfd ensembles**. Visualization and Computer Graphics, IEEE Transactions on, 19(12):2743–2752, 2013.

[Hunter2007] HUNTER, J. D.. **Matplotlib: A 2D Graphics Environment**. Computing in Science & Engineering, 9(3):90–95, 2007.

[Joia2011] JOIA, P.; PAULOVICH, F. V.; COIMBRA, D.; CUMINATO, J. A. ; NONATO, L. G.. **Local Affine Multidimensional Projection**. IEEE Transactions on Visualization and Computer Graphics, 17(12):2563–2571, dec 2011.

[Kawas2014] KAWAS, B.; KOC, A.; LAUMANNS, M.; LEE, C.; MARINESCU, R.; MEVISSEN, M.; TAHERI, N.; VAN DEN HEEVER, S. ; VERAGO, R..

**Unified framework and toolkit for commerce optimization under uncertainty**. IBM Journal of Research and Development, 58(5/6):12–1, 2014.

[Keim2008] KEIM, D. A.; MANSMANN, F.; OELKE, D. ; ZIEGLER, H.. **Visual analytics: Combining automated discovery with interactive visualizations**. In: PROCEEDINGS OF THE 11TH INTERNATIONAL CONFERENCE ON DISCOVERY SCIENCE, DS '08, p. 2–14, Berlin, Heidelberg, 2008. Springer-Verlag.

[Kohlhammer2011] KOHLHAMMER, J.; KEIM, D.; POHL, M.; SANTUCCI, G. ; ANDRIENKO, G.. **Solving problems with visual analytics**. In: PROCEEDINGS OF THE 2ND EUROPEAN FUTURE TECHNOLOGIES CONFERENCE AND EXHIBITION 2011 (FET 11), p. 117–120, 2011.

[Kruskal1978] KRUSKAL, J. B.; WISH, M.. **Multidimensional Scaling**, volumen 31. 1978.

[Lee2010] LEE, H. K. H.; TADDY, M. ; GRAY, G. A.. **Selection of a Representative Sample**. Journal of Classification, 27(1):41–53, mar 2010.

[MacEachren2012] MACEACHREN, A. M.; ROTH, R. E.; O'BRIEN, J.; LI, B.; SWINGLEY, D. ; GAHEGAN, M.. **Visual Semiotics & Uncertainty Visualization: An Empirical Study**. IEEE Transactions on Visualization and Computer Graphics, 18(12):2496–2505, dec 2012.

[Meira2016] MEIRA, L. A.; COELHO, G. P.; SANTOS, A. A. S. ; SCHIOZER, D. J.. **Selection of Representative Models for Decision Analysis Under Uncertainty**. Computers & Geosciences, 88:67–82, mar 2016.

[Numpy2011] VAN DER WALT, S.; COLBERT, S. C. ; VAROQUAUX, G.. **The NumPy Array: A Structure for Efficient Numerical Computation**. Computing in Science & Engineering, 13(2):22–30, mar 2011.

[Park2007] PARK, K.; CAERS, J.. **History Matching in Low-Dimensional Connectivity-Vector Space**. In: EAGE CONFERENCE ON PETROLEUM GEOSTATISTICS, número September 2007, p. 10 – 14, 2007.

[Park2016] PARK, H.; BELLAMY, M. A. ; BASOLE, R. C.. **Visual analytics for supply network management: System design and evaluation**. Decision Support Systems, 91:89 – 102, 2016.

[Paulovich2008] PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R. ; LEVKOWITZ, H.. **Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping**. IEEE Transactions on Visualization and Computer Graphics, 14(3):564–575, 2008.

[Paulovich2010PLP] PAULOVICH, F. V.; SILVA, C. T. ; NONATO, L. G.. **Two-Phase Mapping for Projecting Massive Data Sets**. IEEE Transactions on Visualization and Computer Graphics, 16(6):1281–1290, nov 2010.

[Phadke2012] PHADKE, M. N.; PINTO, L.; ALABI, O.; HARTER, J.; TAYLOR II, R. M.; WU, X.; PETERSEN, H.; BASS, S. A. ; HEALEY, C. G.. **Exploring ensemble visualization**. In: IS&T/SPIE ELECTRONIC IMAGING, p. 82940B–82940B. International Society for Optics and Photonics, 2012.

[Rahimian2018] RAHIMIAN, H.; BAYRAKSAN, G. ; HOMEM-DE MELLO, T.. **Identifying effective scenarios in distributionally robust stochastic programs with total variation distance**. Mathematical Programming, jan 2018.

[Sahaf2016] SAHAF, Z.; HAMDI, H.; MAURER, F.; NGHIEM, L. ; SOUSA, M. C.. **Clustering of geological models for reservoir simulation studies in a visual analytics framework**. In: 78TH EAGE CONFERENCE AND EXHIBITION 2016, 2016.

[Sanyal2010] SANYAL, J.; SONG ZHANG; DYER, J.; MERCER, A.; AMBURN, P. ; MOORHEAD, R. J.. **Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty**. IEEE Transactions on Visualization and Computer Graphics, 16(6):1421–1430, nov 2010.

[Sarma2013] SARMA, P.; CHEN, W. H. ; XIE, J.. **Selecting Representative Models From a Large Set of Models**. In: SPE RESERVOIR SIMULATION SYMPOSIUM. Society of Petroleum Engineers, feb 2013.

[Schardong2018] SCHARDONG, G. G.; RODRIGUES, A. M.; BARBOSA, S. D. ; LOPES, H.. **Visual interactive support for selecting scenarios from time-series ensembles**. Decision Support Systems, 113:99–107, sep 2018.

[Scheidt2009a] SCHEIDT, C.; CAERS, J.. **Representing spatial uncertainty using distances and kernels**. Mathematical Geosciences, 41(4):397–419, 2009.

[Scheidt2009b] SCHEIDT, C.; CAERS, J.. **Uncertainty Quantification in Reservoir Performance Using Distances and Kernel Methods–Application to a West Africa Deepwater Turbidite Reservoir**. SPE Journal, 14(04):680–692, dec 2009.

[ScikitLearn2012] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; LOUPPE, G.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M. ; DUCHESNAY, É.. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 12, jan 2012.

[Scipy2001] JONES, E.; OLIPHANT, T.; PETERSON, P. ; OTHERS. **SciPy: Open source scientific tools for Python**, 2001.

[Shirangi2016] SHIRANGI, M. G.; DURLOFSKY, L. J.. **A general method to select representative models for decision making and optimization under uncertainty**. Computers & Geosciences, 96:109–123, nov 2016.

[Suzuki2006] SUZUKI, S.; CAERS, J. K.. **History Matching With an Uncertain Geological Scenario**. In: SPE ANNUAL TECHNICAL CONFERENCE AND EXHIBITION. Society of Petroleum Engineers, apr 2006.

[Szafir2018] SZAFIR, D. A.. **The good, the bad, and the biased: Five ways visualizations can mislead (and how to fix them)**. Interactions, 25(4):26–33, June 2018.

[Tufte1990] TUFTE, E.. **Envisioning Information**. Graphics Press, 1990.

[Wallner2000] WALLNER, J.; KRASAUSKAS, R. ; POTTMANN, H.. **Error propagation in geometric constructions**. CAD Computer Aided Design, 2000.

[Wang2014] WANG, Y.; TANG, S.; ZHANG, Y.-D.; LI, J.-T. ; WANG, D.. **Representative selection based on sparse modeling**. Neurocomputing, 139:423–431, 2014.

[Waser2014] WASER, J.; KONEV, A.; SADRANSKY, B.; HORVÁTH, Z.; RIBIČIÇ, H.; CARNECKY, R.; KLUDING, P. ; SCHINDLER, B.. **Many plans: Multidimensional ensembles for visual decision support in flood management**. Computer Graphics Forum, 33(3):281–290, 2014.

[Wong2013] WONG, C.; OLIVEIRA, M. C. F. ; MINGHIM, R.. **Multidimensional Projections to Explore Time-Varying Multivariate Volume Data**. In: GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), 2013 26TH CONFERENCE ON, p. 107–114, 2013.

[Yang2007] YANG, Y.; WEBB, G. I.; CERQUIDES, J.; KORB, K. B.; BOUGHTON, J. ; TING, K. M.. **To select or to weigh: A comparative study of linear combination schemes for superparent-one-dependence estimators**. Knowledge and Data Engineering, IEEE Transactions on, 19(12):1652–1665, 2007.

# A
# User profile questionnaire for prototype evaluation sessions

1. Personal information

   (a) Name

   _____

   (b) E-mail

   _____

   (c) Education degree

   _____

   (d) Course semester (just for students)

   _____

   (e) Occupation

   _____

2. Please mark below your knowledge about the following subjects:

   (a) Division of a sample by percentiles: P10, P50, P80 ...

   ◯ I do not know

   ◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)

   ◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)

   ◯ I know well (I do not apply often, but I would not need to revise the concepts if I had to apply them)

   ◯ I am a specialist (I apply these concepts frequently)

   (b) Analysis of trends and patterns in time series

   ◯ I do not know

   ◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)

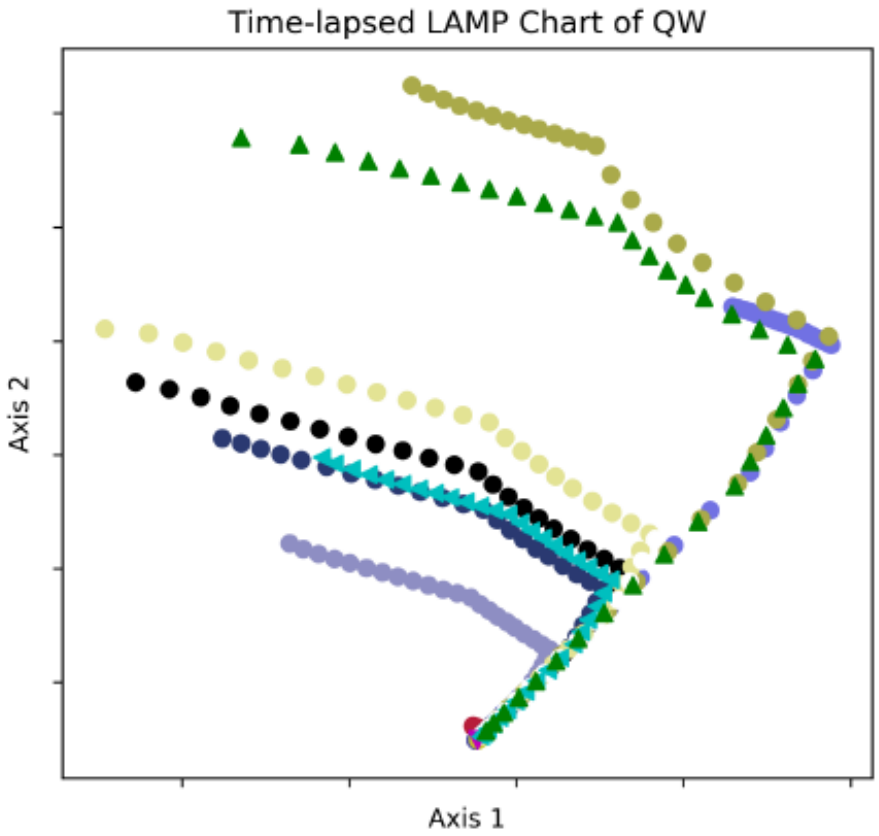
   ◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)

   ◯ I know well (I do not apply often, but I would not need to revise

the concepts if I had to apply them)

◯ I am a specialist (I apply these concepts frequently)

(c) Projection chart (Time-lapsed LAMP chart)




◯ I do not know

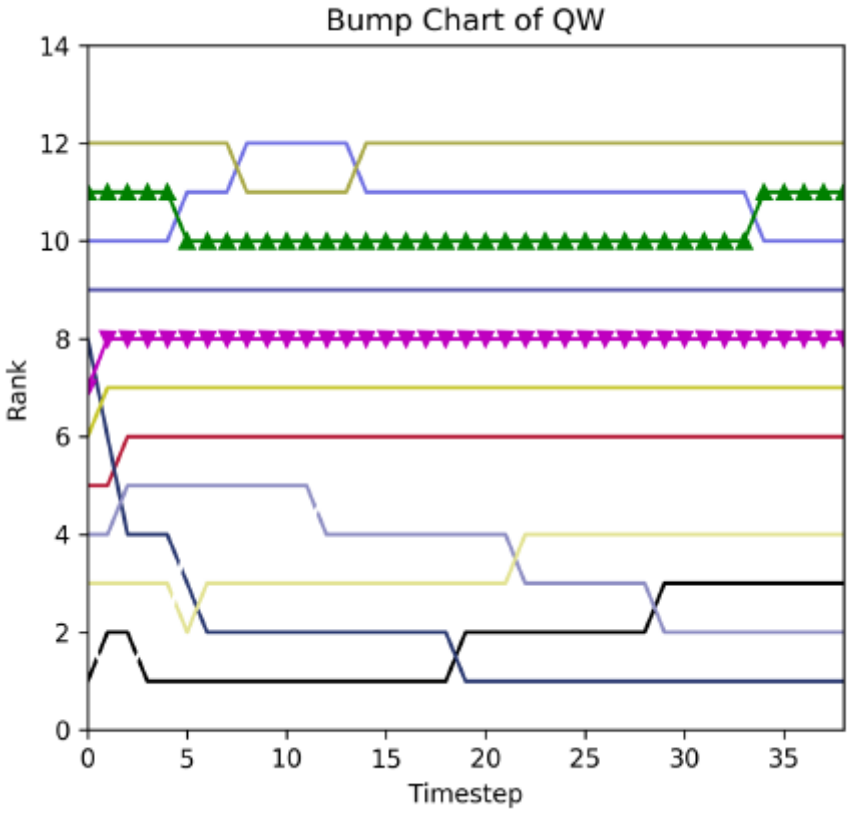
◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)

◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)

◯ I know well (I do not apply often, but I would not need to revise the concepts if I had to apply them)

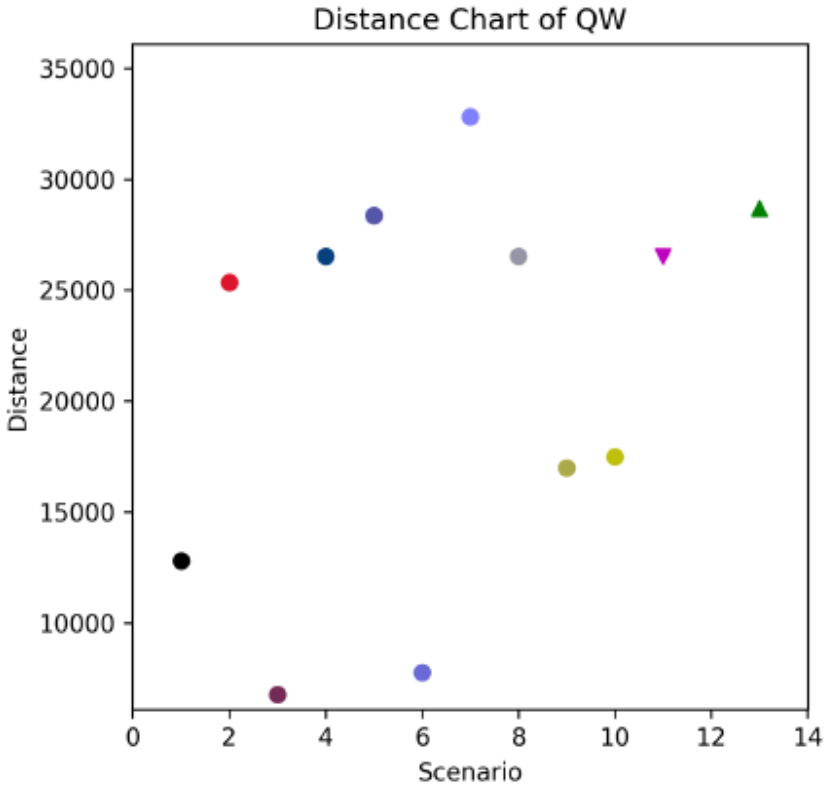
◯ I am a specialist (I apply these concepts frequently)

(d) Ranking chart (Bump chart)



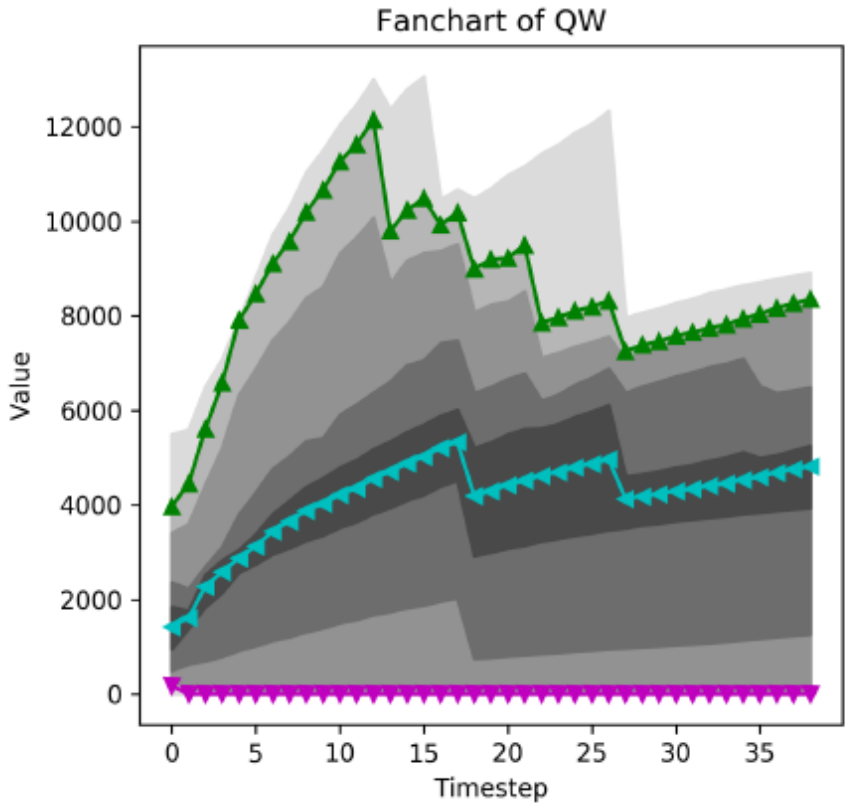

- ◯ I do not know
- ◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)
- ◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)
- ◯ I know well (I do not apply often, but I would not need to revise the concepts if I had to apply them)
- ◯ I am a specialist (I apply these concepts frequently)

(e) Distance chart




- ◯ I do not know
- ◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)
- ◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)
- ◯ I know well (I do not apply often, but I would not need to revise the concepts if I had to apply them)
- ◯ I am a specialist (I apply these concepts frequently)

(f) Fanchart




◯ I do not know

◯ I know little (I have learned these concepts at some point, but may have to learn again if I have to apply them)

◯ I have average knowledge (I may have to revise one concept or another if I have to apply it)

◯ I know well (I do not apply often, but I would not need to revise the concepts if I had to apply them)

◯ I am a specialist (I apply these concepts frequently)