

# Definition of Desirable Scenarios

Deliverable D4.2 – WP4 – PU




# Definition of Desirable Scenarios

Work package 4, Deliverable D4.2

**Please refer to this report as follows:**

M. Zach, C. Rudloff, M. Sawas (2019). Definition of Desirable Scenarios. Deliverable D4.2 of the H2020 project LEVITATE.

Project details:	
<b>Project start date:</b>	01/12/2018
<b>Duration:</b>	36 months
<b>Project name:</b>	LEVITATE – Societal Level Impacts of Connected and Automated Vehicles
<b>Coordinator:</b>	Pete Thomas, Prof – Prof. of Road & Vehicle Safety Loughborough University Ashby Road, LE11 3TU Loughborough, United Kingdom
	The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824361.

Deliverable details:	
<b>Version:</b>	Final
<b>Dissemination level:</b>	PU (Public)
<b>Due date:</b>	30/09/2019
<b>Submission date:</b>	29/11/2019

**Lead contractor for this deliverable:**

Martin Zach – AIT Austrian Institute of Technology

<b>Report Author(s):</b>	Zach, M., Rudloff, C., Sawas, M. (AIT Austrian Institute of Technology) Austria
--------------------------	---

## Revision history

Date	Version	Reviewer	Description
15/10/2019	Preliminary draft 1	(distributed to LEVITATE list)	Review round 1
18/11/2019	Draft for review	Anastasios Dragomanovits (NTUA) Evita Papazikou (Loughborough University)	Review round 2
28/11/2019	Final draft		Updates after 2 <sup>nd</sup> review
29/11/2019	Final report	Camellia Hayes (Loughborough University, English language reviewer)	Sanity check before submitting
29/11/2019	Final deliverable	Pete Thomas – Loughborough University → EC	

## Legal Disclaimer

All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user, therefore, uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and INEA has no liability in respect of this document, which is merely representing the authors' view.

# Table of contents

<b>Executive Summary .....</b>	<b>1</b>
<b>1 Introduction .....</b>	<b>3</b>
<b>1.1 LEVITATE .....</b>	<b>3</b>
<b>1.2 Work package 4 and Deliverable 4.2 within LEVITATE .....</b>	<b>3</b>
<b>1.3 Organization of the deliverable .....</b>	<b>4</b>
<b>2 Background and related work.....</b>	<b>5</b>
<b>2.1 Correlations between indicators within - and across - the dimensions considered in LEVITATE .....</b>	<b>5</b>
<b>2.2 Defining visions for backcasting approaches.....</b>	<b>7</b>
<b>2.3 Application of statistical &amp; machine learning techniques .....</b>	<b>8</b>
<b>3 Data sources and inputs for data analysis .....</b>	<b>11</b>
<b>3.1 Open Data sources considered .....</b>	<b>11</b>
<b>3.2 Inputs from the Stakeholder Reference Group .....</b>	<b>14</b>
<b>3.3 Organisation of data according to dimensions &amp; goals, levels and time .....</b>	<b>16</b>
<b>4 Examples of Visions.....</b>	<b>28</b>
<b>4.1 City of Vienna.....</b>	<b>28</b>
<b>4.2 Greater Manchester.....</b>	<b>33</b>
<b>4.3 "Vision Zero".....</b>	<b>35</b>
<b>5 Selected approach for defining desirable visions.....</b>	<b>37</b>
<b>5.1 Detailed description of selected indicators and geo-entities .....</b>	<b>37</b>
<b>5.2 Statistical approach (PCA based data imputation) .....</b>	<b>43</b>
<b>5.3 Collaborative filtering approach .....</b>	<b>44</b>
<b>6 Visualisations and interpretation of results.....</b>	<b>48</b>
<b>6.1 Results of statistical analysis .....</b>	<b>48</b>
6.1.1 Similarity of indicators.....	48
6.1.2 Similarity of geo-entities.....	51
6.1.3 Development over time .....	51
6.1.4 Identifying of visions .....	54
<b>6.2 Results of collaborative filtering .....</b>	<b>56</b>
6.2.1 Similarity of indicators.....	56
6.2.2 Similarity of geo-entities.....	57
6.2.3 Development over time and identifying of visions .....	58

6.3	Common interpretation of results .....	60
7	Conclusions and outlook.....	62
7.1	Identifying visions in City strategies .....	62
7.2	Identification of feasible transformation paths .....	62
7.3	Backcasting process.....	63
References	.....	65
Appendix	.....	68
	Used Terminology.....	68

# List of Figures

Figure	Page
Figure 1: Correlations and relationships between LEVITATE policy goals (examples)	7
Figure 2: Eurostat: Availability of relevant data by geographical level	12
Figure 3: Correlation Matrix of the indicators for all years available.	50
Figure 4: (a) Variances of the principal components of the imputed data. (b) Plot of the first two principal components of the PCA on the imputed data.	51
Figure 5: Importance of the variables for the first two PCs.	52
Figure 6: First two components of the PCA on the twice imputed data set.	53
Figure 7: First five PCs for the data of the city of Vienna.	55
Figure 8: Visualisation of LEVITATE indicators and four LEVITATE dimensions in embedding space (first two primary components). Geo-Entities are also shown as yellow dots.	57
Figure 9: Illustration of clustering for geo-entities in two-dimensional projection of the embedding space (top: PCA1, PCA2; bottom: PCA1, PCA3). Colour legend is explained in text.	59
Figure 10: Development over time (1960 – 2010) for geo-entity Vienna and Visions 2030 and 2050.	60

# List of Tables

Table	Page
Table 1: Eurostat data sets and relevant indicators	12
Table 2: Inputs from the Stakeholder Reference Group	14
Table 3: Selected data sets and indicators – structured along dimensions and goals	17
Table 4: Selected data sets and indicators – structured along data sources	23
Table 5: City of Vienna - Impact targets and Goals	28
Table 6: City of Vienna - Indicators with their specified target values	31
Table 7: City of Vienna - selection of Smart City targets, matching LEVITATE goals & indicators	32
Table 8: Greater Manchester Vision - Impact targets and Goals	34
Table 9: Vision Zero: Definition of proposed Indicators	35
Table 10: LEVITATE Indicators Mapping for Vision Zero	36
Table 11 Indicators used for the statistical analysis of the indicator data	38
Table 12: Mapping of LEVITATE goals and indicators to quantitative targets defining a vision	63

# Executive Summary

---

The main goal of this deliverable is the identification of desirable visions, based on the quantified policy goals and the corresponding indicator framework that has been developed in deliverable D4.1. Challenging questions in this process are:

- How to prioritize different goals across the four dimensions considered in LEVITATE (Safety, Society, Environment and Economy)
- Which relationships between different goals can be identified? (Are they supporting each other or are they conflicting?)

The analyses performed within this deliverable on available data help to get a better understanding on how to define a vision related to CATS for a city or region in a quantitative way and to describe feasible transformation paths to reach such a vision. That process (defining “feasible paths of interventions”) will be the scope of deliverable D4.3.

A focussed survey of literature regarding relationships and correlations among the policy goals and indicators considered in LEVITATE shows that even on high level, quite complex relationships are revealed, forming a “network” of interactions. A good amount of the correlations between goals is positive (this means goals are supporting each other). For some relationships such simple statements are not possible (because there might be several contradicting causal relations). And finally, some goals are obviously conflicting to a certain extent – mainly prosperity (and related economic indicators) opposed to environmental indicators.

Defining desirable visions is the starting point for the backcasting approach proposed for LEVITATE. Even though only a few examples can be found within the transport domain, the available literature gives support regarding the methodologies that can be applied for (semi-)quantitative backcasting and specification of visions. From statistical perspective, the challenges for the analysis of available data lie primarily in high dimensionality (of indicators considered) and high sparsity in the data set (out of all combinations of indicators, city (geo-entity) and year (time), only a small percentage is available). This situation leads to the selection of following two approaches to be applied: principal component analysis (PCA) with data imputation and collaborative filtering.

During the collection of open data for the indicators defined in LEVITATE, several data sources have been analysed in detail, and the inputs from the Stakeholder Reference Group have been considered. For the final evaluation, data from two open data sources have been considered: European Statistical Office (Eurostat) and World Development Indicators (WDI). These data are organized along dimensions & goals (the indicator framework developed in deliverable D4.1), geographic levels (country / region / city) and time.

Based on these data, a closer analysis of example visions – with focus on CATS and the LEVITATE indicator framework – is performed: for the two Cities (regions) of Vienna and Greater Manchester, and for “Vision Zero” (putting extreme emphasis on the Safety dimension).



The details of the applied approaches – statistical (PCA based with data imputation) and collaborative filtering – are then described in more detail, in particular how the collected data have to be processed, how they are evaluated and what are the restrictions. The goals and expected outcomes are also explained for each approach: analysing how “close” several indicators are to each other (similarity of indicators), further analysing the similarity for geo-entities, investigating the development (evolution) over time, and finally, how to identify a vision that has been specified by means of the LEVITATE indicator framework.

The main results of data analysis are:

- Similarities (i.e. correlations) between indicators are investigated in a systematic way, showing – by and large – consistency between the two selected approaches and with previous results found in the literature. Nevertheless, also a few surprising results are found: For example, hardly any correlation between road deaths and injuries, and if any it even tends to be negative.
- Clustering of geo-entities is quite strong – cities in the same (European) region (in the same decade) show very similar behaviour.
- Development over time (how geo-entities move in indicator space over the decades) is also clearly visible.
- There are several ways how to map and illustrate a concrete vision (based on specific target values for a city or region) with slightly different but consistent results.

Both discussed approaches, on the other hand, have clear limitations and suffer from the high sparsity in the available data set, despite the methods that have been applied. It should also be noted that visualizing the results in a two-dimensional plot can easily be misleading since it is based on further dimensional reduction.

Nevertheless, these results can be considered as a base for the next task in WP4 – the closer analysis of “feasible paths” towards a desired vision. From the results presented in this deliverable, the “structure” of the indicator space, the observed development at present and the “direction” towards the desired vision are the main inputs. This will be combined now with the preliminary results from other work packages (WP3 – CATS impacts and methods for forecasting them, WP5-7 – (sub) use cases and applications, policy interventions to be considered) and with additional inputs from (and dialogues with) the stakeholder reference group, in the actual backcasting process, outlining feasible paths of intervention.

---

# 1 Introduction

## 1.1 LEVITATE

Societal **Level Impacts** of Connected and **Automated Vehicles** (LEVITATE) is a European Commission supported Horizon 2020 project with the objective to prepare a new impact assessment framework to enable policymakers to manage the introduction of connected and automated transport systems, maximise the benefits and utilise the technologies to achieve societal objectives.

Specifically, LEVITATE has four key objectives:

1. To incorporate the methods within a **new web-based policy support tool** to enable city and other authorities to forecast impacts of connected and automated transport systems (CATS) on urban areas. The methods developed within LEVITATE will be available within a toolbox allowing the impact of measures to be assessed individually. A Decision Support System will enable users to apply backcasting methods to identify the sequences of CATS measures that will result in their desired policy objectives.
2. To develop a range of **forecasting and backcasting** scenarios and baseline conditions relating to the deployment of one or more mobility technologies that will be used as the basis of impact assessments and forecasts. These will cover three primary use cases – automated urban shuttle, passenger cars and freight services.
3. To establish a **multi-disciplinary methodology** to assess the short, medium and long-term impacts of CATS on mobility, safety, environment, society and other impact areas. Several quantitative indicators will be identified for each impact type.
4. To apply the methods and **forecast the impact of CATS** over the short, medium and long-term for a range of use cases, operational design domains and environments and an **extensive range of mobility, environmental, safety, economic and societal indicators**. A series of case studies will be conducted to validate the methodologies and to demonstrate the system.

## 1.2 Work package 4 and Deliverable 4.2 within LEVITATE

The objective of work package 4 is to develop target scenarios and feasible paths to reach them with interventions concerning automated vehicles, contributing mainly to the second LEVITATE objective. The main steps are:

- Research of national/European policy goals in the impact dimensions
- Definition and description of goals and visions<sup>1</sup> of cities and other stakeholders for short, medium and long-term.
- Applying the resulting CATS impacts (from WP3) and data available from the cities to define targets.

---

<sup>1</sup> The term “visions” is used here instead of the term “scenarios” that has been used in the project proposal. Refer also to relevant part of terminology agreed in the project, given in the Appendix (Used Terminology).

- Using backcasting methodologies to define feasible paths to reach the stakeholders' goals with special consideration to automated vehicles
- Definition of forecasting scenarios and desired outputs for the consolidation of the different use-cases

Deliverable 4.2 has as its main goal the definition and description of targets and visions (desirable futures), based on the policy goals and indicators described in Deliverable 4.1, on short-term and longer-term strategy documents from Cities, and on statistical data available for the selected indicators on different geographical levels.

This document describes how to identify visions, analyses the constraints and possible conflicts between targets, and sets the ground for the more detailed investigations of feasible transformation paths to approach these visions (which will be described in Deliverable D4.3).

## 1.3 Organization of the deliverable

This deliverable is organized as follows:

Chapter 2 briefly describes related work and approaches that can be considered as basis for further investigations: studies on correlations and dependencies between the indicators relevant for LEVITATE, backcasting approaches with similar context defining desirable futures, and basic statistical and data science methods for the evaluation of available data.

Chapter 3 documents the data sources and used input data in detail, structuring them according to the indicator framework developed for LEVITATE. This is followed by a discussion of a few example visions in more detail in Chapter 4, where the definition is based on this indicator framework.

Chapter 5 describes the methodologies applied, distinguishing between more traditional statistical methods involving principal component analysis (PCA) and data imputation, and – as an alternative approach – collaborative filtering (which is mostly known from recommender systems<sup>2</sup>). Even if the selected approaches themselves are not extremely sophisticated, the preparation of the available data is a quite complex process, making this chapter rather technical.

The results and visualisations are then presented and discussed in chapter 6, including the illustration of the example visions "Vienna 2030" and "Vienna 2050". Conclusions and outlook to the next steps in WP4 are presented in Chapter 7.

---

<sup>2</sup> Recommender systems are tools for interacting with large and complex information spaces. They support users in finding items of interest, based on a user profile derived from historical data.

## 2 Background and related work

### 2.1 Correlations between indicators within - and across - the dimensions considered in LEVITATE

In the first workshop of the Stakeholder Reference Group, where the definition of indicators relevant for LEVITATE was discussed, the following question arose: "Which indicators could be contradictory (see [1])?"

It is one of the starting points for this deliverable to analyse the literature in terms of observed correlations between the indicators relevant to this project.

Obviously for some indicators there will be positive correlation, for example if the way of calculating them is similar (as in the case of many economic indicators). For other pairs of indicators, even if they belong to the same LEVITATE dimension, an estimation in advance is simply not possible. In general, the LEVITATE indicators have been chosen in such a way that redundancy between indicators is avoided.

Since the research topic of possible correlations between indicators is huge by definition, there is no attempt to cover this in a comprising way. A few examples of investigated correlations or dependencies are listed below.

(Zahabi, et al., 2012) demonstrate in their study that "the built environment (BE) attributes are statistically significant (10% increase in density, transit accessibility and land-use mix, results in 3.5%, 5.8% and 2.5% reduction in GHG respectively)[...]" and consequently highlight the relationship between settlement structure and greenhouse gas emissions.

A paper by (Rode, et al., 2014) investigates how transport and urban form contribute to accessibility in cities. It confirms a correlation between population density and carbon emissions. "For example, at similar wealth levels, sprawling Atlanta produced six times more transport-related carbon emissions than relatively compact Barcelona (ATM, 2013; D'Onofrio 2014; LSE Cities 2014). This finding aligns with analysis conducted for 30 cities in China, which showed that compact cities have higher CO2 efficiency, particularly as a result of supporting non-motorised transport (Liu, Chen et al. 2012)."

In a literature Review the correlation between public transport accessibility and job opportunities was explained. The mentioned correlation has attracted the researcher's attention in the literature. "Researchers have revealed several impact and correlation of provision of public transport accessibility to the environment and daily life which would have a noticeable impact on public health and other aspects of public daily life (M. A. Saif, 2018)."

The Organisation for Economic Cooperation and Development (OECD) published the report "How Was Life?: Global well-being since 1820"(2014) (Zanden, et al., 2014) which contains the research of correlations between different dimensions and GDP per capita. In relation to quality of environment it concludes that "Per capita emissions of CO2 and SO2 show, as expected, a positive correlation with GDP per capita, so richer countries do have more emissions. The correlation with CO2 emissions shows a rising trend, which

seems to end after 1970. However, the correlation between GDP and SO<sub>2</sub> is becoming less strong and less significant over time."

A study of the National Research and Development Institute for Industrial Ecology (2015) (Danciulescu, et al., 2015) in Bucharest shows a linkage between noise levels and air pollution. The data was measured and processed with a statistical analysis program. "The results obtained in the tests presented in this paper reveal good and very good correlation between noise level and concentration of NO<sub>2</sub>, SO<sub>2</sub> and CO in the air, in two areas of Bucharest where the most important common source of pollution is traffic."

(Najaf, et al., 2018) investigate the connection of urban form and traffic safety. "In addition to spatial variation in employment and urban density that have significant direct effect on traffic safety, improving transportation network connectivity and increasing the supply of public transit facilities and upper-level transport infrastructures can decrease traffic fatalities indirectly, through encouraging the use of non-driving transport modes. It is estimated that a 10% increase in urban density as well as a 10% increase in even spatial distribution of employment can reduce the rate of fatal crashes by >15%, on average."

A paper by (Mohan, et al., 2017) analyses the impact of urban street structure on traffic safety in U.S. cities and concludes "that (a) higher number of junctions per road length was significantly associated with lower motor vehicle crashes and pedestrian mortality rates, and, (b) more roads of any kind is associated with increased fatality rates while an additional km of main arterial road is associated with higher total fatalities and is statistically significant. The higher the ratio of highways and main arterial roads is against non-arterial roads, the higher the fatality." It shows a correlation of efficient settlement structure respectively use of public space and protection of human life.

(Ingvardson & Nielsen, 2019) compare the relationship of satisfaction and public transport use in six European cities. "The study found that travel satisfaction is positively related to (i) accessibility measures, e.g. extent of network coverage, travel speed and service frequency, (ii) perceived costs, e.g. reasonable ticket prices, and (iii) norms, i.e. perceived societal and environmental importance of public transport."

(Elvik, 2000) summarizes estimates of road accidents costs for the national economy. "On the average, the total costs of road accidents, including an economic valuation of lost quality of life, were estimated to about 2.5% of the gross national product. Excluding the valuation of lost quality of life, road accident costs on the average amounted to 1.3% of the gross national product. When valuation of lost quality of life is included, costs ranged from 0.5 to 5.7% of GNP. When valuation of lost quality of life is disregarded, costs ranged from 0.3 to 2.8% of GNP." The paper therefore demonstrates the relationship between protection of human life and prosperity.

In the figure below correlations and relationships that have been found between LEVITATE policy goals are outlined. Red and green lines demonstrate positive and negative correlations respectively, found in the literature. Grey lines symbolize relationships that are neither expected to be clearly positive nor expected to be clearly negative. Dotted lines show correlations of indicators that are – by definition – closely related to each other, without giving specific evidence found in literature. As noted before, these should be considered as examples rather than an exhaustive list.

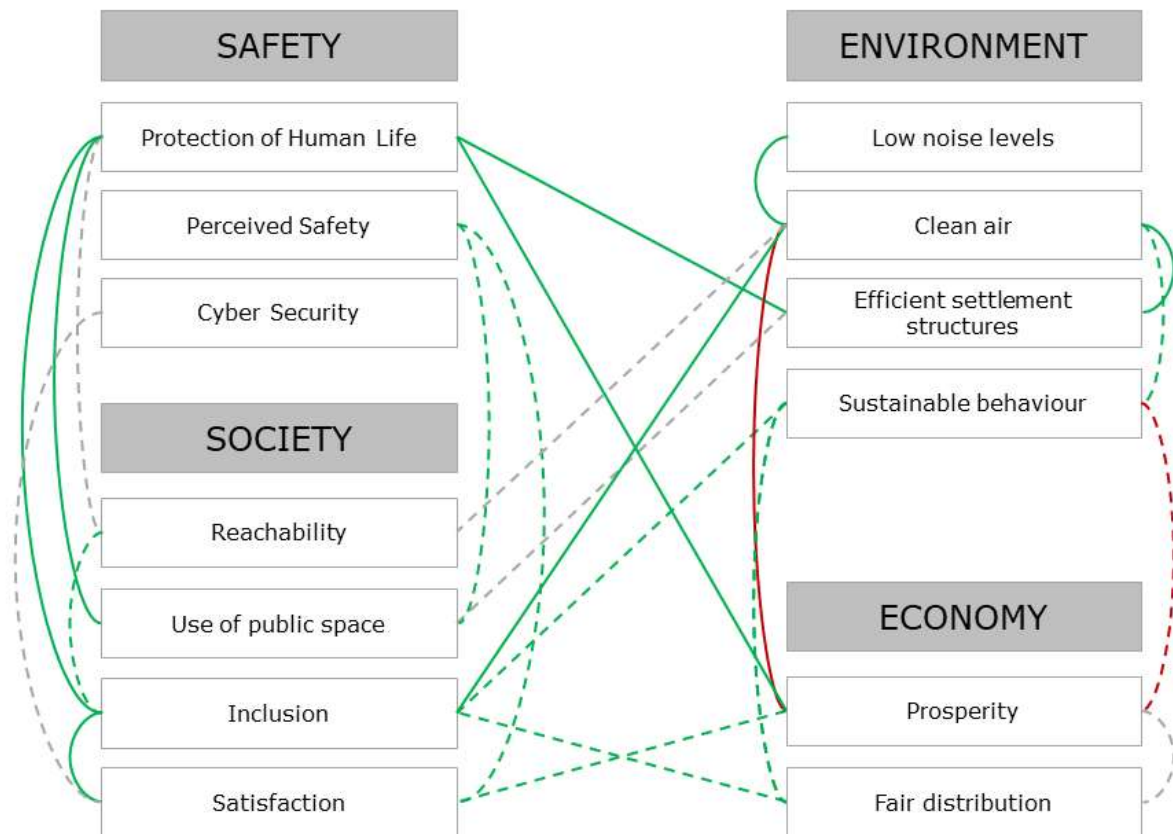


Figure 1: Correlations and relationships between LEVITATE policy goals (examples)

## 2.2 Defining visions for backcasting approaches

The backcasting approach proposed for LEVITATE tries to address the following question: “What measures need to be taken (and when) in order to realise a specific (quantified) objective set for a specific year?” The corresponding task is to estimate the contributions of various programmes or measures towards realising the targets, thus putting together a package of actions policymakers can take to ensure that the objective is realised. As an example, a city might consider policy objectives of preventing an increase in travel distances of fossil fuel vehicles, increasing walking and cycling, and, at the same time, reducing the number of traffic injuries. Realising all objectives at the same time might be quite challenging, especially if these objectives are “conflicting” (showing “negative correlation”).

The set of objectives for a specific year in the future is referred to as Vision in this Deliverable. The precise definition as agreed in the LEVITATE Terminology Guide (refer also to Appendix) is: “Description of a future situation defined by a bundle of vision characteristics and dedicated at a specific point in time.” It should be noted that the term “Vision” is used instead of the term “desired future scenario” that was used in the project proposal, in order to avoid any confusions with simulation scenarios in LEVITATE context.



Investigating related work in which quantitative backcasting approaches have been applied, several methodologies can be found to envision a desirable future. For example, in one approach, a discrete choice experiment was conducted to elicit future ecosystem services demand (Brunner, 2016).

In an introductory paper for presentation of several backcasting studies for sustainability the authors describe the attempts to map a systemic multidimensional concept like sustainability to one or a few simplified quantitative targets (Vergragt, 2011). They also point out that in backcasting “both vision development and pathway development encompass processes of higher order learning, in which participants learn not only about preferable futures and their contradictions, but also about the present, [...] and about how to improve the future vision to make it more appealing and resilient.”

Backcasting studies specific to the field of CATS – which would be most relevant for LEVITATE – are naturally quite rare up to now; one recently completed Austrian research project “System Scenarios Automated Driving in Personal Mobility” – SAFiP (SAFiP, 2019) has at least touched the backcasting aspect for this topic, focusing on two fixed target indicators: traffic volume of motorized individual traffic, and greenhouse gas (CO<sub>2</sub>) emissions. The impact of various influencing factors and additional measures (policy interventions) was then studied by means of a system dynamics approach, resulting in a recommended bundle of interventions in order to reach the defined goals.

Finally, it should also be mentioned that participatory visioning is frequently used to support the identification of desirable visions. Experiences of such processes are described in detail in (Soria-Lara, 2017) for example in the field of transport, considering involvement of stakeholders. It should be noted that in LEVITATE such a participatory approach has been chosen from the beginning, considering the outputs from the first Stakeholder Reference Group workshop particularly for defining the key indicators and targets, their prioritization and interrelationship, as described already in (Zach, 2019).

## **2.3 Application of statistical & machine learning techniques**

The indicator data in the Levitate project are high-dimensional with more than 20 indicators and as such it is difficult to find underlying connections in the data. In particular it is difficult to visualize the structure in the data. As a result, dimension reduction techniques need to be applied to the data. There are many different techniques for such a dimension reduction. Classical techniques like principal component analysis (PCA) (Pearson, 1901) or factor analysis can be applied to rotate the underlying coordinate systems by using linear combinations of the indicators such that the resulting coordinate system (principal components) are chosen to maximise the variance in the data along the first components. Extensions and other methodologies were developed over the years, most coming from the area machine-learning. These include methodologies like KernelPCA (Schölkopf B., 1997), Autoencoders (Hinton, 2006) or collaborative filtering (Herlocker, et al., 1999). The first two methodologies, however, were not designed to deal with high sparsity in the data set.

The first step in the data analysis step was to actually collect available data. While a lot of data can be found as Open Government Data (OGD), data are not easy to bring together on a common geographical basis. Data are available for different points in time and on different geographic scales ranging from city level to country level. This makes it difficult to consolidate available data. As a result, the dataset has many missing data points. Either data are not available on the required geographic level or for all the years and hence time series analysis methods are difficult to apply.

As a result, as mentioned above, simple correlation analysis and dimension reduction techniques cannot be applied directly to the data. The only possible classical way to study the raw data is to apply pairwise comparisons of indicator vectors, but missing data makes this method unreliable and might lead to misleading results, however, in particular for variables with higher data density, pairwise comparison of the variables gives a first valuable insight into the structure of the data as long as caution is taken for variables with a larger number of missing values.

There are many algorithms for the imputation of missing data (Asif, 2013). One way to deal with the multitude of missing data are to impute it using algorithms like that suggested by Josse and Husson 2012 (Josse, 2012) where the missing values are initialised either with mean or random values and are iteratively reassigned using principal component analysis to learn about connections within the existing data. There are restrictions to the imputation algorithm. Years where no initial data or data without variation exists cannot be imputed. Hence the resulting data still has the data points for the different indicators existing for different years. These indicators however are imputed for all geographical regions.

Nevertheless, the data can be used for first analysis. For example, a correlation analysis and PCA can be applied to the imputed data to get information about the structure in the data, correlations in time and different indicators.

In recent years, additional techniques have been developed and applied in order to manage the high sparsity of input data, across a variety of domains. One example is Collaborative filtering (CF) (Herlocker, et al., 1999), a technique primarily used by recommender systems, making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). These predictions can then be used to recommend certain items (e.g. books, movies) to a user (user-item system). It has been recognized that this technique can be used in a more general way, filtering for information or patterns using collaboration among multiple agents, viewpoints, data sources, etc.

While Collaborative filtering, to a certain extent, avoids the necessity to impute missing data, the challenge of data sparsity remains. In the context of indicators and geographical regions this means: if for one region no data (or only data for very few indicators) is available, it will not be possible to make any meaningful prediction for other indicators. This is also referred to as cold start problem.

Matrix factorization (Koren, et al., 2009) is a class of collaborative filtering algorithms used in recent years, decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices. This approach fosters an easy interpretation, as the dimensions of the corresponding lower dimensional space can be considered as latent factors, characterizing both users and items in the same space (transferred to our context: both geographical regions and indicators). In the end, it



comes again to a dimensionality reduction, and by applying PCA on top of that, one can visualize the results in simple 2D plots.

The techniques of collaborative filtering and matrix factorization are also establishing the relationship to a class of methods that has been widely adopted in machine learning, mainly in the Natural Language Processing (NLP) domain: Embeddings (in NLP, words or phrases from the vocabulary are mapped to vectors of real numbers, referred to as word embeddings). Conceptually this involves a mathematical embedding from a space with many dimensions per word to a continuous vector space with a much lower dimension. A generalization of this concept beyond NLP is Entity Embeddings of Categorical Variables (Cheng Guo, 2016). By mapping similar values close to each other in the embedding space this approach reveals the intrinsic properties of the categorical variables (again we might consider here geographical regions and indicators).

# 3 Data sources and inputs for data analysis

## 3.1 Open Data sources considered

This chapter provides a brief description of data sources used for the process of constructing visions of liveable futures for cities. These visions will be based on the statistical analysis of policy goal indicators with the aim to disclose consistent as well as contradicting goals. Achieving sound and reliable results, which will in turn form the basis for providing expedient backcasting functionalities in the Policy Support Tool, is strongly dependent on the amount and quality of available data.

The data collection focused on the following key points:

- time series of data (preferably larger period)
- short description of the relevant indicators / datasets (validation of relationship to LEVITATE indicators defined in (Zach, 2019))
- priority for selection of indicators according to geographical level, in following order: Cities - Regions - Countries
- consideration of the Stakeholder Reference Group (SRG) inputs and answers to request for data

Furthermore, additional indicators were included in the scheme that can be clearly assigned to the four main LEVITATE dimensions of safety, society, environment and economy.

The data are organized as follows:

- Data source
- Filename (download as Excel Table)
- Indicator code
- Geo code

The two main sources of data for the studies are European Statistical Office (Eurostat) and World Development Indicators (WDI).

### **Eurostat**

Eurostat is an important source of open data and contains the world's most comprehensive data on events in different countries. It also gives access to other records contained in the data catalogue (Eurostat, 2019). Eurostat contains huge amount of statistical data, with 4,600 data sets and 14,000 indicators. With this source many important datasets have been identified. In the search, it was possible to use a filter for important features such as Time or GeoCodes.

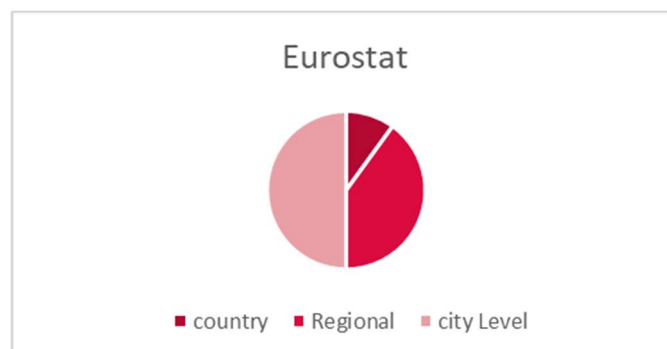


Figure 2: Eurostat: Availability of relevant data by geographical level

The Data collection includes a set of traffic indicators at levels country Level, city level or NUTS 2 and 3 for the road (infrastructure) or traffic accidents. The information can be found in the Eurostat dissemination database (Eurobase) under certain heading like "General and regional statistics / NUTS classification / regional transport statistics" and can be reflected in the theme "Transport / Multimodal data / Regional transport statistics."

Indicators are labelled in a similar way to variables, except that the names end with an I to identify them as indicators.

Every dataset for indicators has an online data code. These online codes used for the analysis are (for the most relevant data sets used):

Table 1: Eurostat data sets and relevant indicators

<b>Online data code</b>	<b>Theme</b>	<b>Relevant indicators</b>	<b>Oldest -, most recent data</b>
tran_r_acci	Transport	Victims in road accidents by NUTS 2 regions	1990, 2017
urb_ctrans	General and regional statistics	Transport- Cities and greater cities	1990, 2019
urb_cenv	General and regional statistics	Environment - cities and greater cities	1990, 2019
urb_clivcon	General and regional statistics	Living conditions - cities and greater cities	1990, 2018
urb_percep	General and regional statistics	perception survey results	2004, 2015

lan_use_ovw	General and regional statistics	Land use overview by NUTS2 regions	2009, 2015
lan_lcv_ovw	General and regional statistics	Land cover overview by NUTS2 regions	2009, 2015
ilc_mddw04	Population and social conditions	Noise from neighbours or from the street by degree of urbanisation	2004, 2015
met_d3dens	Demography statistics	Population density by metropolitan regions	1990, 2016
nama_10r_3gdp or nama_10r_2gdp	Demography statistics	Gross domestic product (GDP) at current market prices by NUTS 3 or 2 regions	2000, 2017
ilc_di12c	Population and social conditions	Gini coefficient of equivalised disposable income before social transfers (pensions excluded from social transfers)	2003, 2018

### WDI (World Development Indicators)

WDI is the primary World Bank collection of world development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates (WDI, 2019).

The database contains 1,600 time series indicators for 217 economies (on country level) and more than 40 country groups (geographical and other groupings), with data for many indicators going back more than 50 years (WDI, 2019).

## 3.2 Inputs from the Stakeholder Reference Group

With the help of Stakeholder Reference Group, information could be provided about potentially available data that will be considered in the project for the best possible results.

A list of the indicators that should be included in the statistical analysis was provided to the stakeholders. For each indicator, it was briefly stated in their answers whether data are available for the city / region and whether they are aware of specific conditions for the collection and use of such data. The open data was provided with a weblink.

Table 2: Inputs from the Stakeholder Reference Group

Di-mension	Indicator	Ut-recht	Bu-da-pest	KiM&SWOV	Lough	Ma-drid	NTUA 2	Wiener Linien	Vi-enna	Greater Manches-ter
<b>Safety</b>	Number of injured per million in-habitants	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Number of fatalities per million in-habitants (per year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Standard-ized survey: subjective rating of (overall) safety	No	No	No	No	Yes	Yes	Yes	Not an-swerd	Not an-swerd
	Number of successful attacks per million trips completed	No	No	No	NO	Yes	No	Yes	Not an-swerd	Not an-swerd
	Number of vulnerabili-ties found (fixed?) (per year)	No	No	NO	No	yes	No	Yes	Not an-swerd	Not an-swerd
<b>Society</b>	Average travel time per day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No an-swer	Not an-swerd
	Number of opportuni-ties / points of interest per 30 minutes per mode of transport	No	No	Not sure	Not sure	No	Yes	No	po-ten-tially	Not an-swerd

Di- mension	Indicator	Ut- recht	Bu- da- pest	KiM& SWOV	Lough	Ma- drid	NTUA 2	Wiener Linien	Vi- enna	Greater Manches- ter
	Lane space (area) per person	Not sure	Yes	Not answerd	Not sure	Yes	Yes	Yes	could be done	Not answerd
	Pedestrian/cycling space (area) per person	Not sure	Yes	Not answerd	Not sure	No	Yes	No	No	Not answerd
	Distance to nearest publicly accessible transport stop (including MaaS)	Yes, but no MaaS included	Yes	Not answerd	Not sure	Yes	No	Yes	could be done	Not answerd
	Affordability/discounts	Not sure	Yes	Not answerd	Not sure	No	Yes	Yes	similar indicator	Not answerd
	Barrier free accessibility	Not sure	Yes	Not answerd	Not sure	No	Yes	Yes	No	Not answerd
	Quality of access restrictions/scoring	Not sure	No	Not answerd	Not sure	Yes	No	No	Not answerd	Not answerd
	Satisfaction with active transport infrastructure (walking / cycling) in neighbourhood	Yes	No	Not answerd	Not sure	No	Yes	In German	Yes	Not answerd
	Satisfaction public transport in neighbourhood	Yes	No	Not answerd	Not sure	No	Yes	In German	Yes	Not answerd
Envi- ron- ment	Standardized survey: subjective rating of main sources of disturbing noise	No	No	Not answerd	Not sure	No	No	Yes	Yes	Not answerd

Di- mension	Indicator	Ut- recht	Bu- da- pest	KiM& SWOV	Lough	Ma- drid	NTUA 2	Wiener Linien	Vi- enna	Greater Manches- ter
	Emissions directly measurable: SO2, PM2,5, PM10, NO2, NO, NOx, CO, O3	Yes	Yes	Not answerd	Yes	Yes	Yes	Yes	No answer	Yes
	Building volume per square kilometre (total and per built-up area)	Maybe	No	Not answerd	Not sure	Yes	No	Partial	Could be done	Not answerd
	Population density	Maybe	Yes	Not answerd	yes	Yes	Yes	Yes	Yes	Not answerd
	Rate of energy consumption per person (total)	Not per person but per living-space	Yes	Not answerd	Some data available	No	Yes	Yes	Not answerd	Not answerd
	Rate of energy consumption per person (transport related)	No	Yes	Not answerd	Some figures available	No	Yes	Partial	No	Not answerd
<b>Economy</b>	Taxable income in relation to purchasing power	Yes	Yes	Not answerd	Not sure	Yes	Yes	Yes	Yes	Not answerd
	GINI index	Yes	Yes	Not answerd	Not sure	Not sure	yes	yes	Not answerd	Not answerd

### 3.3 Organisation of data according to dimensions & goals, levels and time

As described before, the available data are quite heterogeneous, and have to be structured according to several criteria, mainly:

- Mapping to LEVITATE dimensions & goals
- Geographical level and regions considered
- Time

In this section, the list of selected data sets and indicators is documented from two different aspects:

- Structured according to LEVITATE dimensions & goals, following the indicator framework proposed in (Zach, 2019).
- List of used datasets with selected indicators (serving as input for further analysis).

Table 3: Selected data sets and indicators – structured along dimensions and goals

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
<b>Safety</b>	ProtectLife	Injured	NUTS2	Eurostat_tran_r_acci.xls	INJ	Injured_1
		Fatalities	NUTS2	Eurostat_tran_r_acci.xls	KIL	Fatalities_2
		People killed in road accidents per 10000 pop.	City	Eurostat_urban_ctrans.xls	TT1060I	Fatalities_1
		Mortality caused by road traffic injury (per 100,000 people)		WDI	SH.STA.TRAF.P5	Fatalities_3
	PercSafety	You feel safe in the neighbourhood you live in: rarely or never	City	Eurostat_urban_percep.xls	PS3024V	PercSafety_1
		You feel safe in this city: strongly agree	City	Eurostat_urban_percep.xls	PS3290V	PercSafety_2
		You feel safe in this city: strongly disagree	City	Eurostat_urban_percep.xls	PS3293V	PercSafety_3
		Most important in my city: Urban safety	City	Eurostat_urban_percep.xls	PS3211V	PercSafety_4



<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
<b>Society</b>	Reachability	Average time of journey to work	City	Eurostat_urb_ctran.xls	TT1019V	TravelTime_1
	PublicSpace	Transport, communication networks, storage, protective works	NUTS2	Eurostat_lan_use_ovw.xls	LUD6	LaneSpace_1
		Public spaces in this city such as markets, squares, pedestrian areas: very satisfied	City	Eurostat_urb_percep.xls	PS3062V	PedestSpace_1
		Public spaces in this city such as markets, squares, pedestrian areas: not at all satisfied	City	Eurostat_urb_percep.xls	PS3065V	PedestSpace_2
	Inclusion	Cost of a combined monthly ticket (all modes of public transport) for 5	City	Eurostat_urb_ctran.xls	TT1080V	Affordability_1
	Satisfaction	Public transport in the city, for example bus, tram or metro: satisfied	City	Eurostat_urb_percep.xls	PS9112V	SatisfactPubTran_1

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
		Public transport in the city, for example bus, tram or metro: very satisfied	City	Eurostat_urb_percep.xls	PS1012V	SatisfactPubTran_2
		Public transport in the city, for example bus, tram or metro: not at all satisfied	City	Eurostat_urb_percep.xls	PS1015V	SatisfactPubTran_3
	SettleStruct	Built	NUTS2	Eurostat_lan_lcv_ovw.xls	LCA1	BuildingVol_1
		Buildings with 1 to 3 floors	NUTS2	Eurostat_lan_lcv_ovw.xls	LCA11	BuildingVol_2
		Buildings with more than 3 floors	NUTS2	Eurostat_lan_lcv_ovw.xls	LCA12	BuildingVol_3
		Population density by metropolitan regions		Eurostat_met_d3dens.xls	met_d3dens	PopDens_1
		Persons per square kilometre	NUTS3	Eurostat_demo_r_d3dens.xls	PER_KM2	PopDens_2
	LowNoise	In this city, noise is a big problem: strongly agree	City	Eurostat_urb_percep.xls	PS2062V	LowNoise_1
		In this city, noise is a big problem: strongly disagree	City	Eurostat_urb_percep.xls	PS2065V	LowNoise_2
		The noise level in the city: very satisfied	City	Eurostat_urb_percep.xls	PS3270V	LowNoise_3
<b>Environment</b>						

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
		The noise level in the city: not at all satisfied	City	Eurostat_urb_percep.xls	PS3273V	LowNoise_4
		The noise level in the city: satisfied	City	Eurostat_urb_percep.xls	PS9370V	LowNoise_5
		Total	City	Eurostat_urb_percep.xls	TOTAL	LowNoise_6
		Proportion of residents exposed to road traffic noise >65 dB(A) at day time	City	Eurostat_urb_cenv.xls	EN2033I	LowNoise_7
		Proportion of residents exposed to road traffic noise >55 dB(A) at night time	City	Eurostat_urb_cenv.xls	EN2035I	LowNoise_8
	CleanAir	In this city, air pollution is a big problem: strongly agree	City	Eurostat_urb_percep.xls	PS2052V	CleanAir_1
		In this city, air pollution is a big problem: strongly disagree	City	Eurostat_urb_percep.xls	PS2055V	CleanAir_2
		Total	Country	Eurostat_ilmddw04.xls	TOTAL	LowNoise_6
		Number of days ozone O <sub>3</sub> concentrations exceed 120 µg/m <sup>3</sup>	City	Eurostat_urb_cenv.xls	EN2002V	O3_1

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
		Number of hours nitrogen dioxide NO2 concentrations exceed 200 µg/m <sup>3</sup>	City	Eurostat_urb_cenv.xls	EN2003V	NO2_1
		Number of days particulate matter PM10 concentrations exceed 50 µg/m <sup>3</sup>	City	Eurostat_urb_cenv.xls	EN2005V	PM10_1
		Accumulated ozone concentration in excess 70 µg/m <sup>3</sup>	City	Eurostat_urb_cenv.xls	EN2025V	O3_2
		Annual average concentration of NO2 (µg/m <sup>3</sup> )	City	Eurostat_urb_cenv.xls	EN2026V	NO2_2
		Annual average concentration of PM10 (µg/m <sup>3</sup> )	City	Eurostat_urb_cenv.xls	EN2027V	PM10_2
		CO2 emissions (metric tons per capita)		WDI	EN.ATM.CO2E.PC	CO2_1
		PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)		WDI	EN.ATM.PM25.MC.M3	PM2.5_1

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
	Settlement	Population density (people per sq. km of land area)		WDI	EN.POP.DNST	PopDens_3
	Sustainable Behaviour	Energy use (kg of oil equivalent per capita)		WDI	EG.USE.PCAP.KG.OE	EnergyConsumTot_1
		CO2 emissions from transport (% of total fuel combustion)		WDI	EN.CO2.TRAN.ZS	EnergyConsumTran_1
<b>Economy</b>	Prosperity	Euro per inhabitant	NUTS3	Eurostat_nama_10r_3gdp.xls	EUR_HAB	Income_3
		Purchasing power standard (PPS) per inhabitant	NUTS3	Eurostat_nama_10r_3gdp.xls	PPS_HAB	Income_4
		Euro per inhabitant	NUTS2	Eurostat_nama_10r_2gdp.xls	EUR_HAB	Income_5
		Purchasing power standard (PPS) per inhabitant	NUTS2	Eurostat_nama_10r_2gdp.xls	PPS_HAB	Income_6
		Euro per inhabitant	City	Eurostat_met_10r_3gdp.xls	EUR_HAB	Income_1
		Purchasing power standard (PPS) per inhabitant	City	Eurostat_met_10r_3gdp.xls	PPS_HAB	Income_2
		Median disposable annual household income	City	Eurostat_urb_clivcon.xls	EC3039V	Income_8

<b>Dimension</b>	<b>Goal</b>	<b>Indicator</b>	<b>Global level</b>	<b>Source</b>	<b>Original Indicator Code</b>	<b>LEVITATE Indicator Code</b>
		Average disposable annual household income	City	Eurostat_urb_clivcon.xls	EC3040V	Income_9
		GNI per capita (constant LCU)		WDI	NY.GNP.PCAP.KN	Income_7
	FairDist	GNI per capita (constant LCU)		Eurostat_ilc_di12c.xls	GINI_HND	GINI_2
		GINI index (World Bank estimate)		WDI	SI.POV.GINI	GINI_1

Table 4: Selected data sets and indicators – structured along data sources

<b>Source</b>	<b>Indicator Code (original)</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LEVITATE Indicator Code</b>	<b>Global level</b>
Eurostat_tran_r_acci.xls	KIL	Killed	Safety	Protect-Life	Fatalities_2	NUTS2
Eurostat_tran_r_acci.xls	INJ	Injured	Safety	Protect-Life	Injured_1	NUTS2
Eurostat_urb_ctrans.xls	TT1019V	Average time of journey to work	Society	Reachability	Travel Time_1	City

<b>Source</b>	<b>Indicator Code (original)</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LEVITATE Indicator Code</b>	<b>Global level</b>
Eurostat_urb_ctran.xls	TT1080V	Cost of combined monthly ticket (all modes of public transport)	Society	Inclusion	Affordability_1	City
Eurostat_urb_ctran.xls	TT1060I	People killed in road accidents per 10000 pop.	Safety	Protect-Life	Fatalities_1	City
Eurostat_lan_use_ovw.xls	LUD6	Transport, communication networks, storage, protective works	Society	Public-Space	LaneSpace_1	NUTS2
Eurostat_urb_percep.xls	PS1012V	Public transport in the city, for example bus, tram or metro: very satisfied	Society	Satisfaction	SatisfactPubTran_2	City
Eurostat_urb_percep.xls	PS1015V	Public transport in the city, for example bus, tram or metro: not at all satisfied	Society	Satisfaction	SatisfactPubTran_3	City
Eurostat_urb_percep.xls	PS2052V	In this city, air pollution is a big problem: strongly agree	Environment	CleanAir	CleanAir_1	City
Eurostat_urb_percep.xls	PS2055V	In this city, air pollution is a big problem: strongly disagree	Environment	CleanAir	CleanAir_2	City
Eurostat_urb_percep.xls	PS2062V	In this city, noise is a big problem: strongly agree	Environment	Low-Noise	LowNoise_1	City
Eurostat_urb_percep.xls	PS2065V	In this city, noise is a big problem: strongly disagree	Environment	Low-Noise	LowNoise_2	City
Eurostat_urb_percep.xls	PS3024V	You feel safe in the neighbourhood you live in: rarely or never	Safety	Perc-Safety	PercSafety_1	City
Eurostat_urb_percep.xls	PS3062V	Public spaces in this city such as markets, squares, ped-	Society	Public-Space	PedestSpace_1	City

<b>Source</b>	<b>Indicator Code (original)</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LEVITATE Indicator Code</b>	<b>Global level</b>
		estrian areas: very satisfied				
Eurostat_urb_percep.xls	PS3065V	Public spaces in this city such as markets, squares, pedestrian areas: not at all satisf.	Society	Public-Space	PedestSpace_2	City
Eurostat_urb_percep.xls	PS3211V	Most important in my city: Urban safety	Safety	Perc-Safety	PercSafety_4	City
Eurostat_urb_percep.xls	PS3270V	The noise level in the city: very satisfied	Environment	Low-Noise	LowNoise_3	City
Eurostat_urb_percep.xls	PS3273V	The noise level in the city: not at all satisfied	Environment	Low-Noise	LowNoise_4	City
Eurostat_urb_percep.xls	PS3290V	You feel safe in this city: strongly agree	Safety	Perc-Safety	PercSafety_2	City
Eurostat_urb_percep.xls	PS3293V	You feel safe in this city: strongly disagree	Safety	Perc-Safety	PercSafety_3	City
Eurostat_urb_percep.xls	PS9112V	Public transport in the city, for example bus, tram or metro: satisfied	Society	Satisfaction	SatisfactPubTran_1	City
Eurostat_urb_percep.xls	PS9370V	The noise level in the city: satisfied	Society	Low-Noise	LowNoise_5	City
Eurostat_ilc_mddw04.xls	TOTAL	Total	Environment	Low-Noise	LowNoise_6	Country
Eurostat_lan_lcv_ovw.xls	LCA1	Built	Society	Settle-Struct	BuildingVol_1	NUTS2
Eurostat_lan_lcv_ovw.xls	LCA11	Buildings with 1 to 3 floors	Society	Settle-Struct	BuildingVol_2	NUTS2
Eurostat_lan_lcv_ovw.xls	LCA12	Buildings with more than 3 floors	Society	Settle-Struct	BuildingVol_3	NUTS2
Eurostat_met_d3dens.xls	met_d3dens	Population density by metropolitan regions	Society	Settle-Struct	PopDens_1	
Eurostat_demo_r_d3dens.xls	PER_KM2	Persons per square kilometre	Society	Settle-Struct	PopDens_2	NUTS3
Eurostat_nama_10r_3gdp.xls	EUR_HAB	Euro per inhabitant	Economy	Prosperity	Income_3	NUTS3
Eurostat_nama_10r_3gdp.xls	PPS_HAB	Purchasing power standard	Economy	Prosperity	Income_4	NUTS3



<b>Source</b>	<b>Indicator Code (original)</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LEVITATE Indicator Code</b>	<b>Global level</b>
		(PPS) per inhabitant				
Eurostat_nama_10r_2gdp.xls	EUR_HAB	Euro per inhabitant	Economy	Prosperity	Income_5	NUTS2
Eurostat_nama_10r_2gdp.xls	PPS_HAB	Purchasing power standard (PPS) per inhabitant	Economy	Prosperity	Income_6	NUTS2
Eurostat_met_10r_3gdp.xls	EUR_HAB	Euro per inhabitant	Economy	Prosperity	Income_1	City
Eurostat_met_10r_3gdp.xls	PPS_HAB	Purchasing power standard (PPS) per inhabitant	Economy	Prosperity	Income_2	City
Eurostat_ilc_di12c.xls	GINI_HND	Gini coefficient (scale from 0 to 100)	Economy	FairDist	GINI_2	Country
Eurostat_urb_cenv.xls	EN2002V	Number of days ozone O3 concentrations exceed 120 µg/m³	Environment	CleanAir	O3_1	City
Eurostat_urb_cenv.xls	EN2003V	Number of hours nitrogen dioxide NO2 concentrations exceed 200 µg/m³	Environment	CleanAir	NO2_1	City
Eurostat_urb_cenv.xls	EN2005V	Number of days particulate matter PM10 concentrations exceed 50 µg/m³	Environment	CleanAir	PM10_1	City
Eurostat_urb_cenv.xls	EN2025V	Accumulated ozone concentration in excess 70 µg/m³	Environment	CleanAir	O3_2	City
Eurostat_urb_cenv.xls	EN2026V	Annual average concentration of NO2 (µg/m³)	Environment	CleanAir	NO2_2	City
Eurostat_urb_cenv.xls	EN2027V	Annual average concentration of PM10 (µg/m³)	Environment	CleanAir	PM10_2	City
Eurostat_urb_cenv.xls	EN2033I	Proportion of residents exposed to road traffic noise >65 dB(A) at day time	Environment	Low-Noise	LowNoise_7	City
Eurostat_urb_cenv.xls	EN2035I	Proportion of residents	Environment	Low-Noise	LowNoise_8	City

<b>Source</b>	<b>Indicator Code (original)</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LEVITATE Indicator Code</b>	<b>Global level</b>
		exposed to road traffic noise >55 dB(A) at night time				
Eurostat_urb_clivcon.xls	EC3039V	Median disposable annual household income	Economy	Prosperity	Income_8	City
Eurostat_urb_clivcon.xls	EC3040V	Average disposable annual household income	Economy	Prosperity	Income_9	City
WDI	EG.USE.PCAP.KG.OE	Energy use (kg of oil equivalent per capita)	Environment	Sustain-Behav	EnergyConsumTot_1	Country
WDI	EN.ATM.CO2E.PC	CO2 emissions (metric tons per capita)	Environment	CleanAir	CO2_1	Country
WDI	EN.ATM.PM25.MC.M3	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	Environment	CleanAir	PM2.5_1	Country
WDI	EN.CO2.TRAN.ZS	CO2 emissions from transport (% of total fuel combustion)	Environment	Sustain-Behav	EnergyConsumTran_1	Country
WDI	EN.POP.DNST	Population density (people per sq. km of land area)	Environment	Settlement	PopDens_3	Country
WDI	NY.GNP.PCAP.KN	GNI per capita (constant LCU)	Economy	Prosperity	Income_7	Country
WDI	SH.STA.TRAF.P5	Mortality caused by road traffic injury (per 100,000 people)	Safety	Protect-Life	Fatalities_3	Country
WDI	SI.POV.GINI	GINI index (World Bank estimate)	Economy	FairDist	GINI_1	Country

## 4 Examples of Visions

The goal of this chapter is to highlight LEVITATE related focus topics in longer-term strategies for a few examples (city level - City of Vienna and Greater Manchester, and European level - EU Road Safety Policy Framework 2021-2030 "Vision Zero") and to define a quantified region in indicator space that represents the *Vision* for a specified point in future.

This activity is based on following main inputs:

- Policy goals and indicators collected and discussed with the cities in a previous phase (documented in (Zach, 2019))
- Additional inputs from the Cities specifying their strategies
- Data sources and input data for relevant indicators (as specified in the previous chapter)

As discussed for the examples in more detail below, two basic strategies can be applied to identify quantified visions:

1. Prioritizing a subset of indicators for which an optimisation is sought for a time point in future (without the need to specify exactly the target values for these indicators) – the vision might then be defined by a (weighted) combination of the selected indicators and specifies the target direction for further development. (Note: Obviously such an approach can only work well if the selected indicators are not in significant conflict with each other; it would therefore be beneficial to select indicators which already show positive correlation.)
2. Defining concrete target values for a certain subset of indicators for a specified time point in future – where the development of other indicators is not specified as constraint but assumed to result from correlations between indicators. (For example, if transport related energy is reduced by 50%, the values for indicators related to air pollution are also likely to improve.)

### 4.1 City of Vienna

The Viennese Urban Mobility Plan, under the "STEP 2025 Urban Development Plan" (Vienna, 2015) sets out the goals of the City of Vienna for a viable transport system of the future. In the section "Objectives and indicators" the following goals and corresponding impact targets are stated – and are mapped to LEVITATE dimensions, goals and indicators in the table below.

Table 5: City of Vienna - Impact targets and Goals

<b>City Goal</b>	<b>Impact Target</b>	<b>Dimension</b>	<b>Goal</b>	<b>Indicator(s)</b>
<b>Fair</b> – Street space is allocated fairly to a variety of users and sustainable mobility must remain	The total sum of spaces for cycling, walking and public transport in all conversion and urban renewal projects is rising.	Society	PublicSpace Inclusion	Pedestrian Space Affordability

affordable for all.				
<b>Healthy</b> – The share of active mobility in every-day life increases; accident-related personal injuries decline.	The share of people in the Viennese population who are actively in motion for 30 minutes daily as they run their daily errands is to rise from 23% in 2013 to 30% in 2025. The number of traffic casualties and persons injured in traffic accidents declines further.	Environment  Safety	SustainBehaviour  ProtectLife	EnergyConsumption Transport  Injured Fatalities
<b>Compact</b> – Distances covered between work, home, errands and leisure time activities are as short as possible.	The share of trips done on foot or by bike to shop for supplies or accompany someone as well as distances covered for leisure time activities will increase from 38.8% in 2013 to 45% in 2025.	Society  Environment	Reachability  Settlement	TravelTime  Population density
<b>Eco-Friendly</b> – Mobility causes as little pollution as possible, the share of eco-mobility in the trips made in Vienna and its environs is rising.	Modal split changes for the Viennese will be reflected in a move away from 72%:28% in 2013 to 80% of eco-mobility and 20% of car traffic by 2025. Traffic in Vienna will shift to a modal split with a much large share of eco-mobility.	Environment	CleanAir LowNoise SustainBehaviour	(all) (all) EnergyConsumption Transport
<b>Robust</b> – Mobility is as reliable and crisis-proof as possible. Mobility should be possible without necessarily owning a means of transport.	The CO2 emissions caused by transport in the Vienna road network (according to the EMIKAT definition) will decline by about 20%, from roughly 2.1 million tons/year in 2010 to about 1.7 million tons/year in 2025. The public transport system	Environment  Society	SustainBehaviour  Reachability Satisfaction	EnergyConsumption Transport  (all) (all)

	remains very reliable. Bicycle availability rises: By 2025 80% of all households should have a bike at their disposal and 40% of the population should be able to reach a bike sharing station within a maximum reach of 300 meters. By 2025, 50% of the population should have a car sharing location within a maximum distance of 500 meters from their homes.			
<b>Efficient</b> – Resources are used in a more efficient way – helped by innovative technologies and processes.	Absolute final energy consumption of the Vienna transport system (according to the EMIKAT definition) will decline by about 20% to around 7.3 TWh by 2025, compared with roughly 9.1 TWh in 2010.	Environment	SustainBehaviour	EnergyConsumption Transport

Further, a series of indicators have been defined (along with qualitative or even quantitative goals for development until 2025) for the following areas (as also documented in (Zach, 2019)):

- Mobility Behaviour
- Mobility Services, reachability and availability of vehicles
- Transport demand, speeds and traffic safety
- Energy and environment

Again, a mapping to LEVITATE indicators is presented for the most relevant indicators along with their specified target values (note that for some indicators quantitative target values are available, for others only qualitative statements development sought are available: rise, decline or maintain level).

Table 6: City of Vienna - Indicators with their specified target values

<b>Indicator</b>	<b>Definition</b>	<b>Most recent value available</b>	<b>Target value (2025)</b>	<b>LEVITATE Indicator(s)</b>
<b>Mobility Behaviour</b>				
Average distances covered [km]	Average distances the Viennese cover in Vienna [km]	2013: 4.1 km	decline	TravelTime
	Share of errands which Viennese population does on foot within walking distances (1 km)	2013: 25.0%	rise	EnergyConsumption Transport
Modal split in passenger transport	Modal split for the Viennese population, referring to the number of trips (eco-mobility:MIT)	2013: 73:27	80:20	EnergyConsumption Transport
Multimodality	Percentage of population using at least two modes of transport within a week	2013: 52%	rise	(relationship to Levitate indicators, but not covered explicitly)
<b>Mobility Services, reachability and availability of vehicles</b>				
Satisfaction with transport	Satisfaction with public transport (school marks 1-5)	2013: 1.70	rise	SatisFactPubTran
Access to public transport stops	Percentage of the population with an underground/suburban train stop located 500 m or less from home or another public transport stop 300 m or less from home	2013: 97.3%	maintain level	(relationship to Levitate indicators, but not covered explicitly)
Degree of motorisation	Passenger cars per 1,000 inhabitants	2014: 386	decline	(relationship to Levitate indicators, but not covered explicitly)
<b>Transport demand, speeds and traffic safety</b>				
Wiener Linien public transport passengers	Passenger numbers on Wiener Linien per year	2013: 900.1 Mio.	rise	(relationship to Levitate indicators, but not covered explicitly)
Average speed of public transport	Average travel speed of tram / bus, rush / evening hours	2013: 15 – 20 km/h	rise	TravelTime
Accidents	Number of traffic casualties per year	2013: 17	decline	Fatalities
	Number of persons injured in traffic accidents per year	2013: 6,979	decline	Injured
<b>Energy and environment</b>				
Energy consumption	Final energy consumption of the transport sector in	2012: 8,647 GWh	7,300 GWh (minus 20%)	EnergyConsumption Transport

	Vienna 1999: 7,474 7.300 per year, adjusted for EMIKAT calculation [GWh]		comp. to 2010)	
CO2 emissions	Traffic-related CO2 emissions in Vienna, according to EMIKAT	2012: 2,062 kt	1,700 kt (minus 20% comp. to 2010)	CO2
Traffic noise	Traffic noise nuisance in close surroundings of home (cumulative, marks 3-5)	2013:29%	decline	LowNoise
PM10 concentration	PM10 limit values exceeded: Number of days when limit value was exceeded (daily mean value >50 g/m <sup>3</sup> ) p.a.	2013: 26	decline	PM10
	PM10 annual mean value mean value	2013: 25 µg/m <sup>3</sup>	decline	PM10
NO2 concentration	NO2 limit values exceeded: Number of half hours when limit value was exceeded (>200 g/m <sup>3</sup> ) p.a.	2013: 0	maintain level	NO2
	NO2 annual mean value mean value	2013: 51 µg/m <sup>3</sup>	decline	NO2

Finally, the targets are aligned with the Vienna Smart City Strategy document of 2019 (Wien, 2019) which presents some quantitative targets for 2030 as well as for 2050. The following table highlights a selection of these targets, matching LEVITATE goals & indicators:

Table 7: City of Vienna - selection of Smart City targets, matching LEVITATE goals & indicators

<b>Description</b>	<b>Target 2030</b>	<b>Target 2050</b>	<b>LEVITATE Goal / Indicator(s)</b>
<b>Quality of Life</b>			
Vienna focuses on <i>social inclusion</i> in its policy design and administrative activities.			Inclusion / Affordability
<b>Resource Conservation</b>			
Vienna reduces its local per capita <i>greenhouse gas emissions</i> by 50% by 2030, and by 85% by 2050 (compared to the baseline year of 2005).	-50%	-85%	CleanAir / (all)
Vienna reduces its local per capita <i>final energy consumption</i> by 30% by 2030, and by 50% by 2050 (compared to the baseline year of 2005).	-30%	-50%	SustainBehaviour / EnergyConsumption Total
<b>Mobility and Transport</b>			

Per capita CO <sub>2</sub> emissions in the transport sector fall by 50% by 2030, and by 100% by 2050.	-50%	-100%	CleanAir / (all)
Per capita final energy consumption in the transport sector falls by 40% by 2030, and by 70% by 2050.	-40%	-70%	SustainBehaviour / EnergyConsumption Transport
The share of journeys in Vienna made by eco-friendly modes of transport, including shared mobility options, rises to 85% by 2030, and to well over 85% by 2050.	85%	> 85%	SustainBehaviour / EnergyConsumption Transport
By 2030, private motor vehicle ownership falls to 250 vehicles per 1,000 inhabitants.	250 / 1,000		(relationship to above mentioned goals, but not covered explicitly)
At least 70% of all journeys in Vienna continue to be short distances of up to 5km, and the majority are made by bike or on foot.			SustainBehaviour / EnergyConsumption Transport
The volume of traffic crossing the municipal boundaries falls by 10% by 2030.	-10%		SustainBehaviour / EnergyConsumption Transport
<b>Economy and Employment</b>			
The incomes and job satisfaction of Viennese citizens constantly increase, while social inequality declines.			Prosperity / Income, FairDistribution / GINI
<b>Environment</b>			
The city's ongoing provision of local green and open spaces for different target groups within the existing urban fabric keeps pace with population growth.			PublicSpace / (all)
In the interests of people's health and well-being, air, water and soil pollution, noise and heat pollution and light pollution are all minimised as far as possible.			CleanAir / (all), LowNoise

## 4.2 Greater Manchester

As outlined in (Zach, 2019), the strategy for the area comprises seven core principles, each of which shall be applied across their transport network:

- Integrated – allow customers to move seamlessly between modes and services
- Inclusive – provide accessible and affordable transport
- Healthy – promote walking and cycling for local trips
- Environmentally responsible – deliver lower emissions, better quality environment
- Reliable – give customers confidence in journey times
- Safe and secure – reduce road accidents and deaths
- Well maintained and resilient – able to withstand unexpected events and weather conditions



Yet, to derive quantified targets from these principles, or even prioritize them, is not a straightforward task.

In the reports “The Greater Manchester transport strategy 2040” (TfGM, 2019) and “5-year environment plan for Greater Manchester” (Manchester, 2019) the goals of the Greater Manchester are set out for a viable transport system of the future. The Table below shows the City Vision and the impact targets. These are mapped to LEVITATE dimensions, goals and indicators.

Table 8: Greater Manchester Vision - Impact targets and Goals

<b>City Vision</b>	<b>Impact Target</b>	<b>Dimension</b>	<b>Goal</b>	<b>Indicator(s)</b>
Reducing CO <sub>2</sub> emissions	The city of Manchester will have reduced CO <sub>2</sub> from 13.6mt in 2014 to 11mt 2020. A robust low carbon pathway to 2050 at which Greater Manchester can become carbon neutral.	Environment	SustainBehaviour CleanAir	EnergyConsumption Transport CleanAir (all)
Increasing use of active travel modes	The daily trips are made by sustainable modes (walking, cycling or public transport) will increase from 39% in 2019 to 50% in 2040. In 2016-18, 56,4% of short journeys (under 2 km) were completed by walking or cycling (GMCA, 2019).	Environment	SustainBehaviour	EnergyConsumption Transport
Replacing fossil-fuelled private vehicles with zero emission (tailpipe) alternatives and bus fleet	Since 2014 the number of plugs in vehicles in Manchester city has been increased (3.3% of new registered cars). The share of fully electric buses will rise to 3.5% in 2013.	Environment	CleanAir LowNoise SustainBehaviour	CleanAir (all)
Reduce roadside NO <sub>2</sub> levels	The annual average roadside NO <sub>2</sub> will be decline from 39 ug per m <sub>3</sub> in 2016 to less than 30 ug per m <sub>3</sub> (GMCA, 2019).	Environment	CleanAir	CleanAir (all)

	Greater Manchester develop Clean Air Plans to bring levels of NO2 on local roads within legal limits as soon as possible.			
Increase road safety	In 2040, Greater Manchester has the ambition to reduce road traffic accidents as close as possible to zero. In recent years, the number of people killed or seriously injured on roads has been increased from 678 people (2016/17) to 788 people (2017/18).	Safety	ProtectLife	Fatalities Injured

### 4.3 "Vision Zero"

Finally, let's consider one example where a vision is defined simply by focussing on one LEVITATE dimension, in this case safety: "Vision Zero". The EU has reaffirmed its ambitious long-term goal, to move close to zero deaths by 2050 (fatalities related to road accidents). In a further recent working paper (EC, 2019), the authors proposed setting of new interim objectives on the way to "Vision Zero" and a range of key performance indicators for road safety (KPIs) at European level, directly related to the prevention of death and serious injury, to provide focus for intervention strategy and delivery. The proposed indicators are listed below.

Table 9: Vision Zero: Definition of proposed Indicators

<b>Indicator</b>		<b>Definition</b>
1	Speed	Percentage of vehicles travelling within the speed limit
2	Safety belt	Percentage of vehicle occupants using the safety belt or child restraint system correctly
3	Protective equipment	Percentage of riders of powered two wheelers and bicycles wearing a protective helmet
4	Alcohol	Percentage of drivers driving within the legal limit for blood alcohol content (BAC)
5	Distraction	Percentage of drivers NOT using a handheld mobile device
6	Vehicle safety	Percentage of new passenger cars with a EuroNCAP safety rating equal or above a predefined threshold*
7	Infrastructure	Percentage of distance driven over roads with a safety rating above an agreed threshold*
8	Post-crash care	Time elapsed in minutes and seconds between the emergency call following a collision resulting in personal injury and the arrival at the scene of the collision of the emergency services

\* Complementary definitions are foreseen for this KPI.

If these indicators are considered in LEVITATE context, it becomes evident that they are mostly not directly related to CATS (and consequently have not been included as LEVITATE indicators, either). Nevertheless, some of these indicators (e.g. Speed, Vehicle safety, Infrastructure) will definitely be impacted by CATS and in this way will also contribute to the higher-level objective of "Vision Zero" – zero fatalities.

Quantitatively this vision might be described by the following:

Table 10: LEVITATE Indicators Mapping for Vision Zero

Indicator description	Target 2050	LEVITATE Indicator
Number of fatalities per million inhabitants (per year)	0 (-100%)	Fatalities
Number of injured per million inhabitants (per year)	Significant decrease (e.g. -80%)	Injured

## 5 Selected approach for defining desirable visions

This chapter describes the methodologies applied in some more detail, distinguishing between more traditional statistical methods involving principal component analysis (PCA) and data imputation, and – as an alternative approach – collaborative filtering (which is mostly known from recommender systems). Even if the selected approaches themselves are not extremely sophisticated and have been applied multiple times in a lot of different domains, the preparation of the available data is a quite complex process, making this chapter rather technical.

The first step is the refined selection and preparation of the available data that have been already documented in chapter 4. Based on these preparations, specific considerations and assumptions for both approaches are described here, before presenting the results of these calculations in chapter 6.

### 5.1 Detailed description of selected indicators and geo-entities

As described earlier, the main challenge for analysing the existing data collected lies in the fragmentation and scarceness: geo-entities on various levels from countries, NUTS-2/3 regions to individual cities (depending on the specific indicator), and availability of yearly values – some from 1960 until the present, and some for one specific year only.

For further analysis the following subset of data will be used:

- **Geo-Entities:** consider only European entities, based on a merging of all considered datasets from Eurostat, where several distinct geo codes might still be consolidated on NUTS-2 or NUTS-3 level at a later stage (e.g. city code, metro region code, NUTS-3 and NUTS-2 code might be mapped to one single entry for the City of Vienna)  
Note: for part of our investigations, also Central Asia has been included, because on the WDI datasets this appears in the same world region as Europe.
- **Indicators:** for statistical methods consider the indicators listed in section 3.3, or – for collaborative filtering calculations explained in section 5.3 – all those indicators for which at least a specific LEVITATE goal can be assigned; for evaluation and interpretation of results focus on the LEVITATE indicators as identified in section 3.3
- **Time Series / Years:** for details and restrictions refer to next subsections

The main idea for the data evaluation is to exploit any hidden relationships between indicators and identify correlation patterns even across geographical levels and decades.

Table 11 summarizes the LEVITATE indicators that were used in the analysis together with their associated LEVITATE dimension. The last column in the table (TargetVal) simply indicates the sign with which an indicator is considered for the evaluations:

- TargetVal = 1 means that a higher value is better (e.g. Purchasing power standard (PPS) per inhabitant, Average disposable annual household income)

- TargetVal = -1 means that a lower value is better (e.g. Gini coefficient (scale from 0 to 100), Annual average concentration of NO<sub>2</sub> (µg/m<sup>3</sup>), Energy use (kg of oil equivalent per capita))

The statements on correlation (versus anti-correlation) consider this sign of contribution. I.e. positive correlation between two indicators always means that optimizing them at same time is possible, where negative correlation means that there is a conflict.

Table 11 Indicators used for the statistical analysis of the indicator data

<b>Source</b>	<b>Indicator-Code</b>	<b>Indicator Name</b>	<b>Dimension</b>	<b>Goal</b>	<b>LevIndicator Code</b>	<b>Target Val</b>
Eurostat_nama_10r_3gdp.xls	EUR_HAB	Euro per inhabitant	Economy	Prosperity	Income_3	1
Eurostat_nama_10r_3gdp.xls	PPS_HAB	Purchasing power standard (PPS) per inhabitant	Economy	Prosperity	Income_4	1
Eurostat_ilc_di12c.xls	GINI_HND	Gini coefficient (scale from 0 to 100)	Economy	Fair Dist	GINI_2	-1
Eurostat_urb_clivcon.xls	EC3039V	Median disposable annual household income	Economy	Prosperity	Income_8	1
Eurostat_urb_clivcon.xls	EC3040V	Average disposable annual household income	Economy	Prosperity	Income_9	1
WDI	NY.GNP.PCAP.KN	GNI per capita (constant LCU)	Economy	Prosperity	Income_7	1
WDI	SI.POV.GINI	GINI index (World Bank estimate)	Economy	Fair Dist	GINI_1	-1
Eurostat_urb_percep.xls	PS2052V	In this city, air pollution is a big problem: strongly agree	Environment	Clean Air	CleanAir_1	-1
Eurostat_urb_percep.xls	PS2055V	In this city, air pollution is a big problem: strongly disagree	Environment	Clean Air	CleanAir_2	1

Eurostat_urb_percep.xls	PS2062V	In this city, noise is a big problem: strongly agree	Environment	Low Noise	LowNoise_1	-1
Eurostat_urb_percep.xls	PS2065V	In this city, noise is a big problem: strongly disagree	Environment	Low Noise	LowNoise_2	1
Eurostat_urb_percep.xls	PS3270V	The noise level in the city: very satisfied	Environment	Low Noise	LowNoise_3	1
Eurostat_urb_percep.xls	PS3273V	The noise level in the city: not at all satisfied	Environment	Low Noise	LowNoise_4	-1
Eurostat_iloc_mddw04.xls	TOTAL	Total	Environment	Low Noise	LowNoise_6	-1
Eurostat_urb_cenv.xls	EN2002V	Number of days ozone O3 concentrations exceed 120 µg/m³	Environment	Clean Air	O3_1	-1
Eurostat_urb_cenv.xls	EN2003V	Number of hours nitrogen dioxide NO2 concentrations exceed 200 µg/m³	Environment	Clean Air	NO2_1	-1
Eurostat_urb_cenv.xls	EN2005V	Number of days particulate matter PM10 concentrations exceed 50 µg/m³	Environment	Clean Air	PM10_1	-1
Eurostat_urb_cenv.xls	EN2025V	Accumulated ozone concentration in excess 70 µg/m³	Environment	Clean Air	O3_2	-1
Eurostat_urb_cenv.xls	EN2026V	Annual average concentration of NO2 (µg/m³)	Environment	Clean Air	NO2_2	-1

Eurostat_urb_cen v.xls	EN2027V	Annual average concentrati on of PM10 (µg/m³)	Envi- ron- ment	Clean Air	PM10_2	-1
Eurostat_urb_cen v.xls	EN2033I	Proportion of residents exposed to road traffic noise >65 dB(A) at day time	Envi- ron- ment	Low Noise	LowNoise_7	-1
Eurostat_urb_cen v.xls	EN2035I	Proportion of residents exposed to road traffic noise >55 dB(A) at night time	Envi- ron- ment	Low Noise	LowNoise_8	-1
WDI	EG.USE.PCAP. KG.OE	Energy use (kg of oil equivalent per capita)	Envi- ron- ment	Sustai n Behav	EnergyConsum Tot_1	-1
WDI	EN.ATM.CO2E. PC	CO2 emissions (metric tons per capita)	Envi- ron- ment	Clean Air	CO2_1	-1
WDI	EN.ATM.PM25. MC.M3	PM2.5 air pollution, mean annual exposure (microgra ms per cubic meter)	Envi- ron- ment	Clean Air	PM2.5_1	-1
WDI	EN.CO2.TRAN. ZS	CO2 emissions from transport (% of total fuel combustion )	Envi- ron- ment	Sustai n Behav	EnergyConsum Tran_1	-1
WDI	EN.POP.DNST	Population density (people per sq. km of land area)	Envi- ron- ment	Settle- ment	PopDens_3	1
Eurostat_tran_r _acci.xls	KIL	Killed	Safety	Protect Life	Fatalities_2	-1
Eurostat_tran_r _acci.xls	INJ	Injured	Safety	Protect Life	Injured_1	-1

Eurostat_urb_ctrans.xls	TT1060I	People killed in road accidents per 10000 pop.	Safety	Protect Life	Fatalities_1	-1
Eurostat_urb_percep.xls	PS3024V	You feel safe in the neighbourhood you live in: rarely or never	Safety	Perc Safety	PercSafety_1	-1
Eurostat_urb_percep.xls	PS3211V	Most important in my city: Urban safety	Safety	Perc Safety	PercSafety_4	-1
Eurostat_urb_percep.xls	PS3290V	You feel safe in this city: strongly agree	Safety	Perc Safety	PercSafety_2	1
Eurostat_urb_percep.xls	PS3293V	You feel safe in this city: strongly disagree	Safety	Perc Safety	PercSafety_3	-1
WDI	SH.STA.TRAF.P5	Mortality caused by road traffic injury (per 100,000 people)	Safety	Protect Life	Fatalities_3	-1
Eurostat_urb_ctrans.xls	TT1019V	Average time of journey to work	Society	Reachability	TravelTime_1	-1
Eurostat_urb_ctrans.xls	TT1080V	Cost of a combined monthly ticket (all modes of public transport) for 5	Society	Inclusion	Affordability_1	-1
Eurostat_lan_use_ovw.xls	LUD6	Transport, communication networks, storage, protective works	Society	Public Space	LaneSpace_1	-1
Eurostat_urb_percep.xls	PS1012V	Public transport in the city, for	Society	Satisfaction	SatisfactPubTran_2	1



		example bus, tram or metro: very satisfied				
Eurostat_urb_percep.xls	PS1015V	Public transport in the city, for example bus, tram or metro: not at all satisfied	Society	Satisfaction	SatisfactPubTran_3	-1
Eurostat_urb_percep.xls	PS3062V	Public spaces in this city such as markets, squares, pedestrian areas: very satisfied	Society	Public Space	PedestSpace_1	1
Eurostat_urb_percep.xls	PS3065V	Public spaces in this city such as markets, squares, pedestrian areas: not at all satisfied	Society	Public Space	PedestSpace_2	-1
Eurostat_urb_percep.xls	PS9112V	Public transport in the city, for example bus, tram or metro: satisfied	Society	Satisfaction	SatisfactPubTran_1	1
Eurostat_urb_percep.xls	PS9370V	The noise level in the city: satisfied	Society	Low Noise	LowNoise_5	1
Eurostat_lan_lcv_ovw.xls	LCA1	Built	Society	Settle-Struct	BuildingVol_1	-1
Eurostat_lan_lcv_ovw.xls	LCA11	Buildings with 1 to 3 floors	Society	Settle-Struct	BuildingVol_2	-1
Eurostat_lan_lcv_ovw.xls	LCA12	Buildings with more than 3 floors	Society	Settle-Struct	BuildingVol_3	1
Eurostat_met_d3dens.xls	met_d3dens	Population density by	Society	Settle-Struct	PopDens_1	1

		metropolitan regions				
Eurostat_demo_r_d3dens.xls	PER_KM2	Persons per square kilometre	Society	Settle-Struct	PopDens_2	1

## 5.2 Statistical approach (PCA based data imputation)

For the analysis of correlations between the selected indicators, a few basic design decisions have to be made, which are further described in this section:

- On which geographical level (i.e. on which granularity of available data) are the correlations analysed?
- How to handle the big amount of missing data?
- How to consider the availability of time series data?

To perform analysis with statistical techniques the data needs to be at a comparable geographical level. Most data are available from European sources, especially data at city level. Hence, the statistical techniques are applied to data on NUTS3 level. For indicators that are available at a higher level, e.g. NUTS2 or country level, the data from the higher level is used for the NUTS3 level data set where available. In addition, several indicators are only available at certain years.

For each indicator, the data was used for all years in which data was available. That resulted in data from years 1960 to 2018 and a total number of 767 data columns. Even with the data from larger geographical regions used for the NUTS3 level, 45,4% of entries in the indicator data were not available making statistical data analysis unreliable. The first step of the statistical analysis of the data was a simple correlation analysis of the indicators. Due to the missing data, the correlation was calculated by calculating the correlation of each of the indicator/year combinations separately, e.g. the correlation for Income\_1 for the year 2014 was compared with the data for indicator Environment\_4 for 2015 if data was available for both indicator/year combinations.

As described in section 2.4, PCA was chosen as a reliable base method for the imputation and analysis of the structure in the data. This approach uses the underlying structure of the data and a dimension reduction to a less high-dimensional data matrix to add values where data was not available for a geographic entity for certain indicators and years. In addition, PCA was applied to learn about the structure of the data and how to use that for visualising city visions. This can be done since PCA offers easy interpretation of the data structure via the factor loadings of the orthogonal transformation of the indicator space. These factor loadings show how closely different indicators are connected.

In a first PCA step the data was studied for connections within different geographical regions. To do that, the data was sorted into a matrix with the NUTS3 regions as rows with one column per indicator and year where data was available. Here, the time dimension was considered together with the indicator dimension. Missing values were imputed using a PCA based algorithm using the first ten PCs for the imputation.

To get a more detailed understanding of the connections of different indicators over space and time, a second PCA was performed. The data was reshaped into a matrix with each indicator as a column and a row per year and geographical entity. So, in this case

the time dimension was considered together with the geo-entity dimension. Since data was not available for all indicators for all years after the first imputation step, a second imputation step was necessary. Again, a PCA based imputation was performed. Due to the much smaller dimension of variables (only the number of indicators compared to all indicator-year combinations for the first imputation step) the first five PCs were used in the imputation algorithm.

Finally, the PCA for the resulting data set was used to analyse the indicators for geographical regions and to analyse the Vision of Vienna compared to historic data. Since the city visions are usually comprised of a subset of all the available indicators, the values for these indicators are taken to be those that the city wants to reach by a certain date. The remaining indicators for the visions are not known. The most likely outcome expected for the vision for the complete indicator set is calculated by again using the underlying structure of the indicator data found in the second PCA. The unknown indicator values for the visions are added by using PCA based imputation again. These values for the visions are compared current values for the indicators and to ones predicted by fitting a linear model to the time series for each indicator. .

The main expected outcomes of the calculations are the following:

- Interpretation with respect to correlations between different indicators within the LEVITATE set and within the indicators for different points in time.
- Interpretation of the closeness of the indicators with respect to geo-entities: Due to the reduction of important dimensions in the PCA, the results of the first principal components offers a good insight if and how indicators within different geographical regions compare and how they develop over time.
- Identifying *visions* for cities by using the PCA results and known visions to estimate the values that need to be reached for all the indicators.

The calculations were performed in the software environment *R* for statistical computing. For the PCA imputation the package *missMDA* was used.

### 5.3 Collaborative filtering approach

As discussed earlier, one further possible group of techniques that can be applied here is Entity Embedding and Collaborative Filtering – as applied successfully in NLP (word vectors) or in recommender systems. Even if the application area in our context is different, the principles are the same.

While some of the challenges described in the previous approach can be avoided by collaborative filtering (like imputation of missing data), several design decisions have to be made also in this approach, which are further described in this section.

On high level, representations in embedding space of words that are closely related and similar to each other (for word vectors) or of users that have very similar interests (in case of recommender systems) should reflect these similarities.

Compared to a typical recommendation system, where we consider Users and Items which have a Rating value within a specified range, the indicator data analysis can be considered in the following way:

- Users: corresponding to Geo-Entities
- Items: corresponding to Indicators

- Rating values: corresponding to individual values for a specific indicator and a specific geo-entity

The basic idea of the collaborative filtering applied here is that even for the scarce data available (for one Indicator only values for a small subset of geo-entities are available, and also the other way around), profiles can be extracted – for indicators and geo-entities at the same time – in form of *latent features* or *latent factors*, which are defined in a multi-dimensional space. Yet, this embedding space has already significantly lower number of dimensions compared to the number of indicators or the number of geo-entities, which corresponds to a dimensionality reduction compared to the original indicator space<sup>3</sup>. Entities (indicators as well as geo-entities) which are located close to each other in the embedding space of latent factors, can be considered as *similar*.

Within the area of collaborative filtering, one of the simplest methods available - Matrix factorization - is applied. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices. This “lower dimensionality” is exactly the number of *latent factors* involved; it has to be determined which number gives the best trade-off between simplification (& saving of computing time) and preserving the essential features of the system (obviously if this number is chosen too small, significant information will be lost).

Note that the time is not yet considered in this scheme. The most straight-forward approach is to link the Year to the Geo-Entity – in this way the geographic differences (between countries, regions or cities) can be analysed together with the development over time. It means, for example, that City A at Year 2015 might have a very similar “profile” (with respect to considered indicators) as City B at Year 2005. A similar approach has also been selected in the PCA based approach described in the last section, for analysing the development over time and identifying visions for cities.

This approach is based on one simplifying assumption, however: The relationship (correlation) of indicators themselves is considered as stable – not depending on time, and not depending on the region in indicator space. An extension of the collaborative filtering approach to calculate the latent factors as a function of time or for clusters in indicator space separately, would be possible but is beyond the scope of this deliverable.

The main expected outcomes of calculations are the following:

- Interpretation with respect to similarity (compatibility) of indicators – within and across LEVITATE dimensions and goals. It would be expected here that indicators within certain dimension (or even more, assigned to the same goal) are closer to each other than indicators belonging to different dimensions.
- Interpretation with respect to geo-entities: Since they can be mapped to the same embedding space as the indicators, it can be analysed how far they are away from the region of ideal values of specific indicators. Similarly, it can also be analysed which geo-entities are similar to each other. It would be expected here that the data points for certain regions (on global level, or also within Europe) and within a certain decade are rather close to each other (forming of clusters).

---

<sup>3</sup> This simplest way of representing categorical data (in original indicator space) is often referred to as *one-hot encoding*.

- Analysing the development over time for specific geo-entities (e.g. in which direction a city is moving).
- Identifying *visions* as regions in embedding space enclosing the (nearly) ideal-value areas for a set of compatible indicators (which is geometrically more effective than determining this region in original indicator space). This will be discussed in more detail in the next section.

Several further actions have been taken to get slightly better results (improving the signal-to-noise ratio):

- Time Series / Years: Several indicators have been measured only in intervals greater than one year (e.g. every 2 years or every 5 years) or even are available in completely irregular intervals only. In addition, for some indicators available on lower geographic level, e.g. Fatalities in a city per year, a high variance can be observed. In order to smooth the data and handle the issue of many missing data points, an average over three consecutive years is taken, and data points for every second year are considered, i.e.
  - Value for 2017: average of values 2016 – 2018
  - Value for 2015: average of values 2014 – 2016
  - etc.
- From the dataset as described above there is still a very weak coupling between geographical levels: For indicators collected at lower levels, there is (mostly) an aggregated value on country level available, but naturally there is no such link available in the other direction (from indicators collected at country level down to regions or cities). This gap is addressed by following logic: If an indicator is available at country level only, the corresponding value (for each country) is taken over for all related lower-level values (cities, regions). This might seem as an over-simplifying assumption, but still infers significant additional information to lower-level geo-entities for which a lot of the considered indicators are not available. For example, the average income or the GINI index of a country are considered also as relevant indicators for the cities of this country (and in general the variance among countries is much larger than the variance between cities).
- The number of *latent factors* has been set to 20, which seems to give reasonable results for the total number of indicators included in the calculations (128).
- In order to make plausibility checks, several “test Indicators” (e.g. a DUMMY indicator, which has a constant value for a certain set of cities) and “test geo-entities” have been induced into the calculations.

The implementation of the described collaborative filtering approach relies on fast.ai (fast.ai, 2019), a deep-learning framework providing a user-friendly library based on PyTorch (PyTorch, 2019). Fast.ai has the mission to make deep learning easier to use for people from all backgrounds. The fastai library simplifies the training of fast and accurate neural networks with modern best practices. It includes “out of the box” support for vision, text, tabular, and collaborative filtering models used here.

The collaborative filtering package used here contains all the necessary functions to quickly train a model for a collaborative filtering task. It offers two possible options:

1. EmbeddingDotBias - Base “dot model” (Matrix factorization) for collaborative filtering: Creates a simple model with Embedding weights and biases for a given number of “users”, a given number of “items” and a given number of latent factors. Predictions in this model can simply be done by taking the dot product of

the embeddings and adding the bias, and then feeding the result to a sigmoid rescaled to a specified range for the “ratings”.

2. EmbeddingNN – creating a full neuronal network of specified size suitable for collaborative filtering.

Only the first option (EmbeddingDotBias - Matrix factorization) has been used for producing the results presented in this Deliverable.

Identification of certain visions, as the examples described in Chapter 4, can be performed by two principle ways in the collaborative filtering approach.

1. Defining “regions” in embedding space that comprise a collection of indicators which are located close to each other (i.e. similar, not in conflict). These regions in embedding space can also be mapped back to regions in indicator space. As a result, the values for all indicators (not only the ones that are part of the selected vision) can be estimated. A problem with this approach, however, is the necessity to explicitly prioritise (weight) the selected indicators.
2. Explicitly setting target values for certain indicators (while others are not fixed), hereby creating a new fictive geo-entity (e.g. Vienna in the year 2050) that is used as additional input for the collaborative filtering. As a result, this additional geo-entity can also be represented in the embedding space and be compared to the current and previous states of the same region / city. Similarly, to the first method, the values for all indicators for the fictive geo-entity can be estimated.

The second option – which is similar to the approach described in the last section – will be illustrated for the example of Vienna in the next chapter.

# 6 Visualisations and interpretation of results

## 6.1 Results of statistical analysis

### 6.1.1 Similarity of indicators

As described in the last chapter, a simple column-wise correlation matrix was calculated first to study the similarity of indicators. Each column in this matrix corresponds to one indicator at a specific year. The data was ordered by indicator dimensions and years, where only years where data was available were considered. To assure meaningful information, the data for the indicators where the goal is to minimise them (e.g. CO<sub>2</sub> emissions) was multiplied by -1. This means that all resulting indicators simply need to be maximised, and positive correlation indicates that corresponding goals would support each other while negative correlation indicates a possible conflict.

The result of the correlation analysis is visualized in Figure 3. The green areas show positive correlations, i.e. the indicators behave similarly, red areas show negative correlations, i.e. if one indicator increases, the other one decreases. The intensity of the colour shows the strength of the correlation. Grey areas indicate that the indicators do not have enough data points to calculate a significant correlation for these indicators. Indicators that have a lot of grey in their rows/columns have only a few data points and hence the correlations with all other indicators have to be considered carefully.

From the blocks along the diagonal it can be seen that most indicators are well correlated over time (with exception of days with NO<sub>x</sub> concentrations above the limit). The most uniform indicator dimension seems to be economy, where there is a positive correlation between most indicators. The other dimensions show indicators both with positive and negative correlations, and also without apparent correlation. This gives an indication that these dimensions will be less likely to be optimised uniformly in a simple fashion and more measures need to be implemented to steer all indicators of these groups into the right directions.

While this approach has to be regarded carefully due to the sparsity of the input data, one can still use it to either confirm expectations regarding correlations or find unexpected correlations that need to be studied in more detail.

Considering the correlations within the dimensions and also in between dimensions some known connections are validated but also some more surprising ones are revealed e.g.:

- Richer regions (countries with higher GDP) tend to have higher green-house gas (GHG)-emissions.
- O<sub>3</sub> and NO<sub>x</sub> emissions tend to have a negative correlation.
- There is hardly any correlation between road deaths and injuries and if any, this tends to be negative.

While the first is a confirmation of expected correlations in the indicators, the latter two need to be looked at in detail. For the negative correlation of O<sub>3</sub> and NO<sub>x</sub>, there are possible explanations from atmospheric chemistry (e.g. (Jacob, 1999) Chapter 12: ...which indicates that O<sub>3</sub> production increases linearly with hydrocarbon concentrations

but varies inversely with NO<sub>x</sub> concentrations. This case is called the hydrocarbon-limited regime because the O<sub>3</sub> production rate is limited by the supply of hydrocarbons. The dependence of O<sub>3</sub> production on NO<sub>x</sub> and hydrocarbons is very different between the two regimes.”). The last one might be due to the fact that speeds in less densely populated regions are higher and hence accidents are more severe, while there are more less severe accidents in regions with lower speeds but a higher mixture of different modes on the streets.



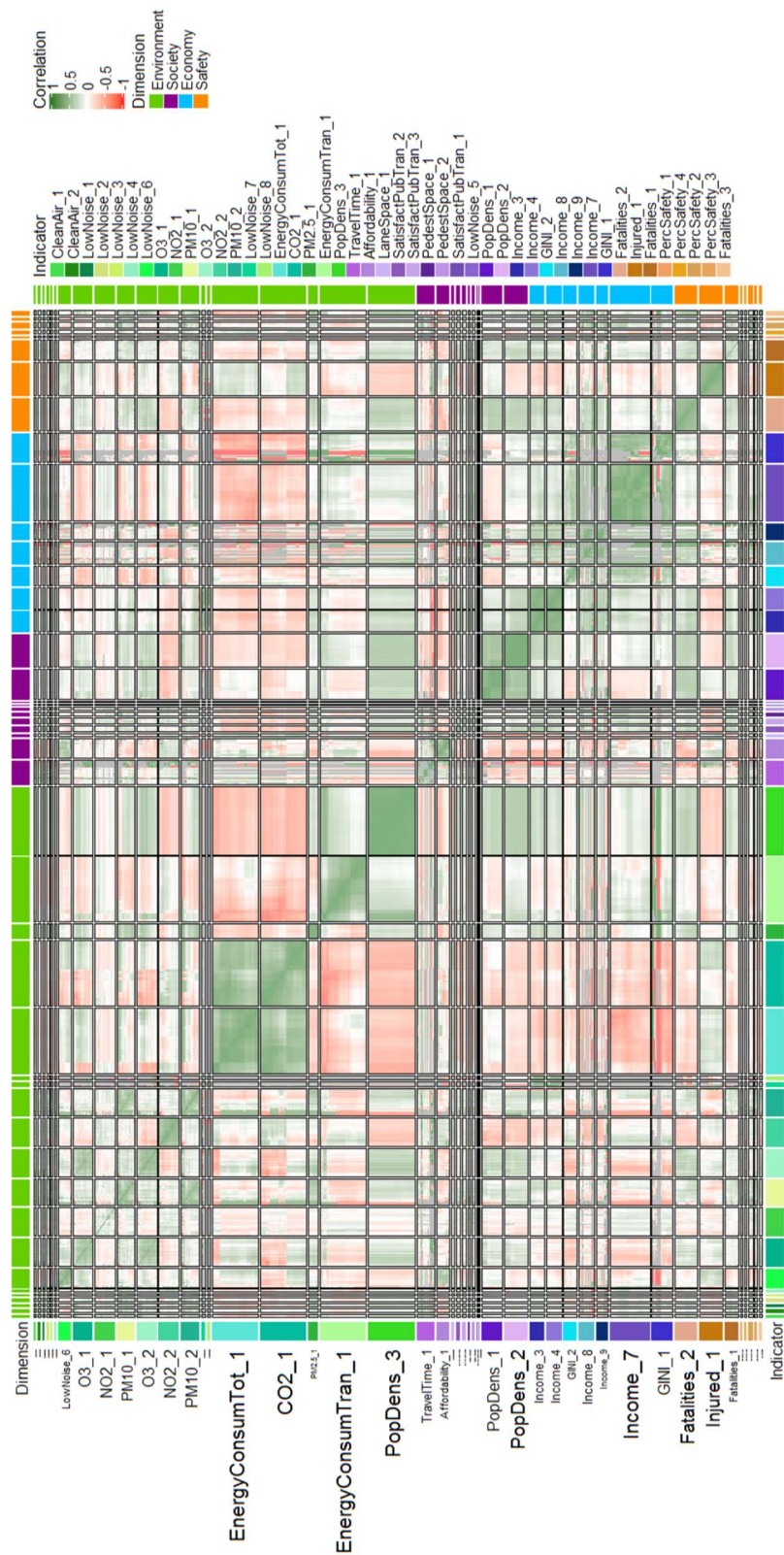


Figure 3: Correlation Matrix of the indicators for all years available. Green areas show positive correlations, red areas negative correlation.

### 6.1.2 Similarity of geo-entities

Due to the sparseness of the indicator matrix, imputation techniques were applied. A PCA-based imputation approach was employed, where a random starting imputation is used and the imputed values were improved using structure in the data with principal component analysis. Due to the large number of columns, the first 10 principal components (PCs) were used in the imputation algorithm. A PCA was then applied to the imputed data set. The results can be seen in Figure 4. The first part of the figure shows that the first components of the imputed data contains a large part of the variance and the variance per PC declines more slowly for the next PCs. The plot shows that the data reduction is feasible due to the fact that a large percentage of the variance is within the first few PCs.

The second plot uses that PCs to visualize the indicator data for the NUTS3 regions grouped by the different geographic regions in Europe (CE- Central Europe, SE - Southern Europe, NE- northern European, EE – Eastern European and GB/UK - British Isles). The first two components which together contain almost 50% of the variance in the data (see Figure 4(a)) of the transformed data are used to visualize the indicator data. One can see that the regions are quite well separated within the first two components and grouped together relatively closely, indicating that the indicators in the different geographical regions behave similarly. While these results seem promising, it is impossible in this approach to see a trend within the data for the different years, and as a result it is difficult to look at the visions of regions within this data set.

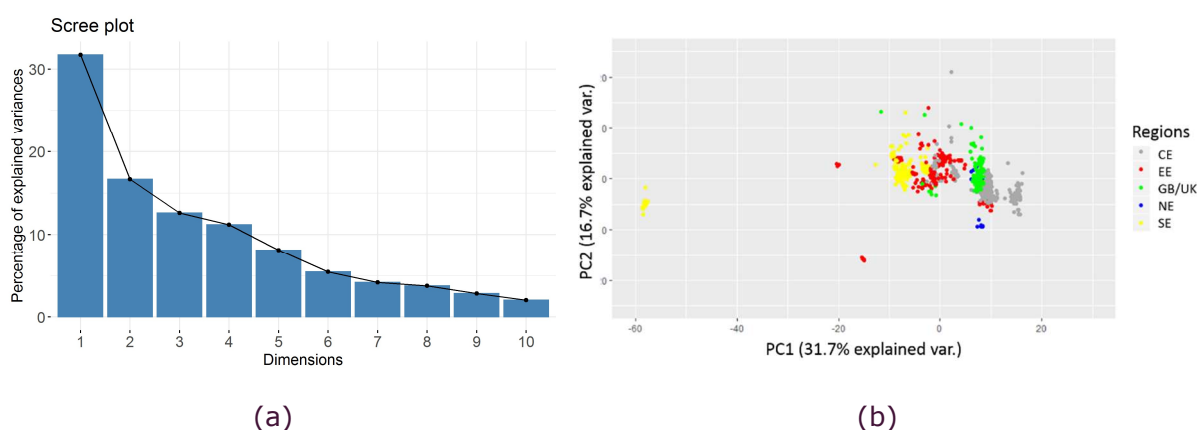


Figure 4: (a) Variances of the principal components of the imputed data. One can see that the main part of the variance of the data is contained in the first few components with the first three components containing more than 50% of the variance in the data. (b) Comparison of the indicators for different geographic region. The comparison is done using the first two PCs. The different geographic regions are plotted in different colours. Points in the same colour are mostly clumped together indicating that the indicators are similar for different the different NUTS3 in the regions.

### 6.1.3 Development over time

As a consequence, a second data set was prepared, that included the development of the indicators over time. The data set was reshaped such that one line of data contained all the 46 chosen indicators for a geographical region and one year. One line per year

(1960-2018) was added to the data for each of the NUTS3 regions. The resulting data contained 87056 lines and 46 columns.

The data again contained a lot of gaps since data was only available for certain years since the first imputation step only included years for each indicator where some data was available. In the new data set each indicator was included for all the years 1960-2018. Overall, 71% of the data in this data-matrix was missing.

A second imputation step was performed, again using a PCA based method for the imputation. Due to the smaller number of columns, only the first five PCs were used within the imputation process.

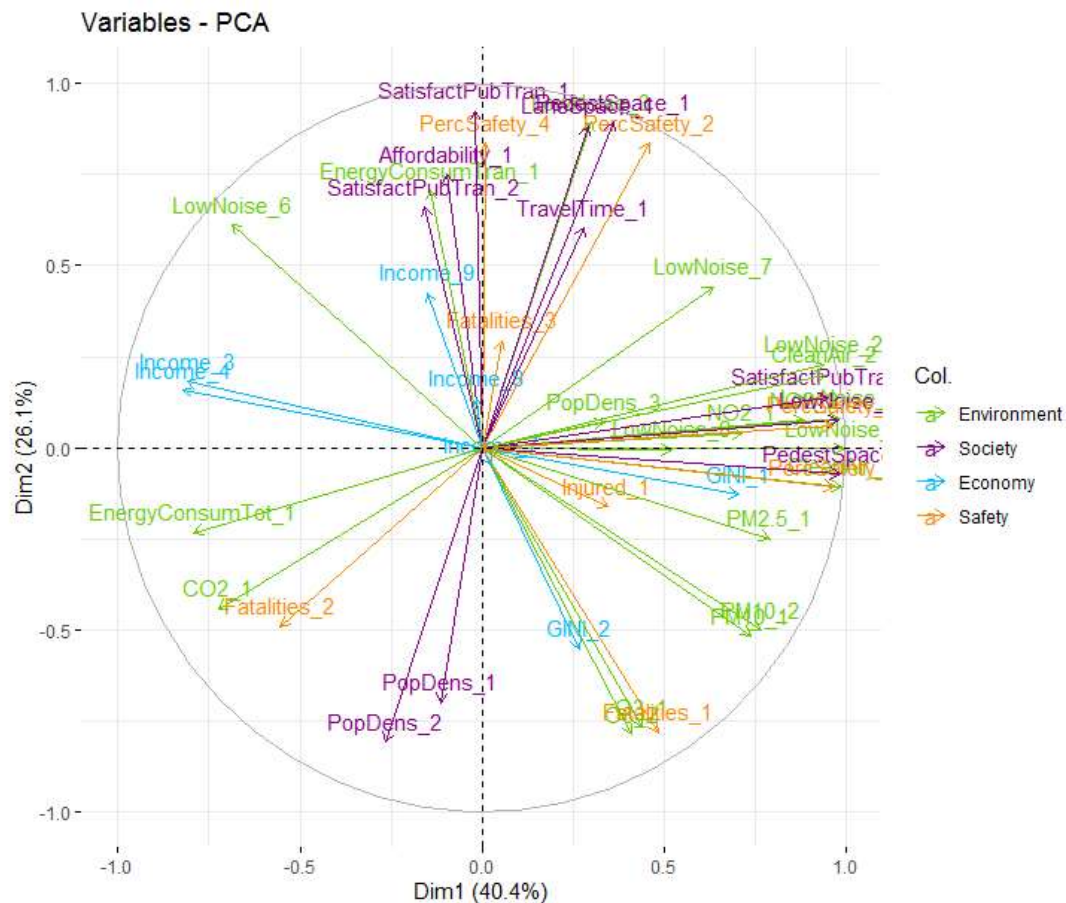


Figure 5: Contribution of the first two PCs to the indicators. The colour of the arrows shows the indicator dimension, the length of the arrow indicates the importance of the indicator for the principal component and indicates the part of the variance of the indicator set that is contained in this indicator. The direction of the arrow gives the importance of the indicator for PC 1 and 2 as well as the sign of the factor loading. Different signs of two factor loadings show that the corresponding indicators have opposing influence on the value of the PC.

Again, a PCA was performed on the imputed data. The result for the first two PCs can be seen in Figure 5. The length of the arrows shows the importance of the indicators in the PCs, the direction the sign of the factor loadings in the PCs. The sign gives an indication of the correlation of the indicators (positive correlation is indicated by a positive sign).

The importance of the indicators shows how much the indicator is considered in the PC. Arrows in the same direction indicate that the indicators behave similarly. In this diagram the importance of each of the indicators for the first two PCs can be seen. The environment and economy dimensions have a strong influence on the first PC whereas society indicators are mostly influencing the second component.

Figure 6 shows the different regions as well as the temporal development of the indicator data within the first five PCs. One can see some temporal development of the PCs for the city of Vienna in particular within the first component where the older values lie more to the left. Looking at the shapes of the dots in the first column of Figure 6, it appears that the points from earlier years are more to the left of the graphs while the newer ones are more to the right. Combining this information with the information of Figure 5, one could deduce that e.g. the environmental dimension might have improved over the years, since the environmental indicators have a strong positive influence on the first PC.

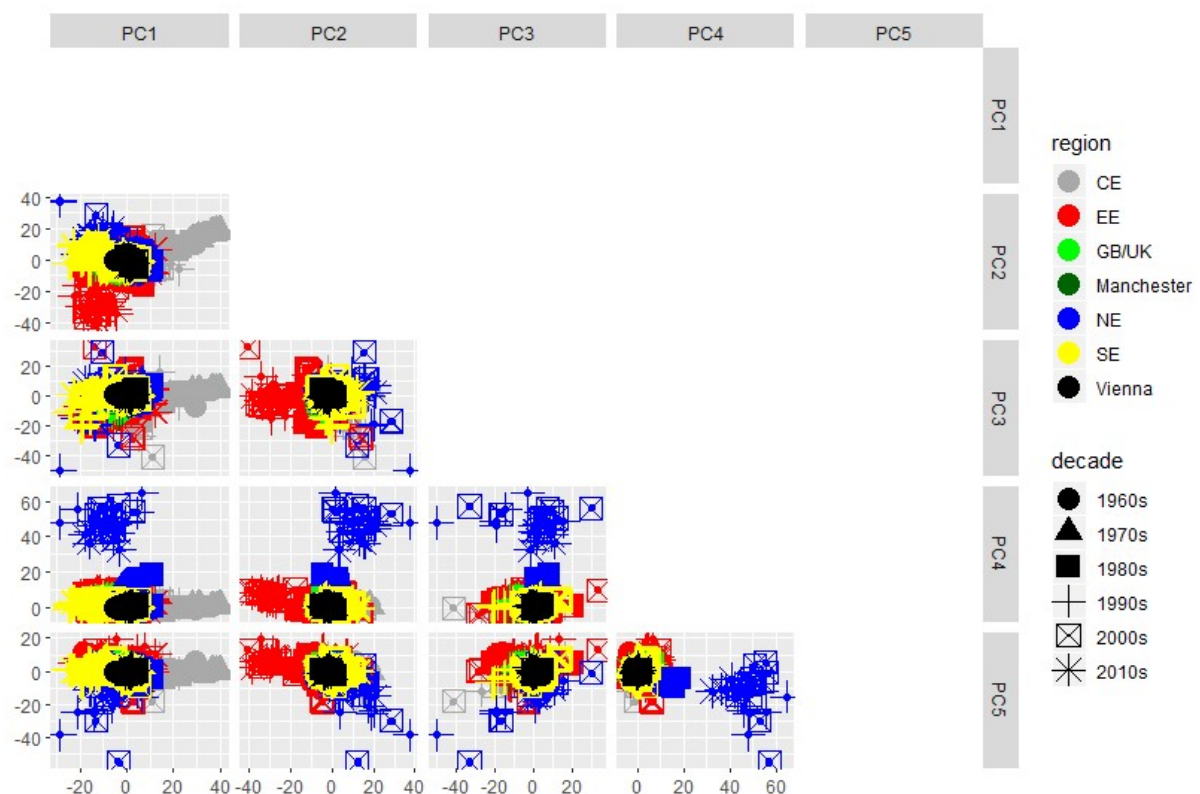


Figure 6: Development of indicators over time - first five components of the PCA on the twice imputed data set. The colours again indicate the different regions in Europe, the shapes the decades that the data originated in. In this plot one can find indications on the development of the PCs over time as well as a regional clustering of the points.



#### 6.1.4 Identifying of visions

Cities set themselves goals for the further improvement of their cities. These improvements can be translated into improvements in the LEVITATE indicator set (for a subset of the indicators and dimensions).

One goal of LEVITATE is to describe paths towards these city visions starting at the current situation. The question that needs to be answered on the path to these visions is whether the indicators are connected strongly and how the final indicator set including the visions of the cities will look like. There are several possibilities to set the values for those indicators which are not explicitly included in the visions themselves.

The PCA gives us a possibility to estimate what the values of unknown indicators within the visions are, since it is based on the interrelations within the data. Consequently, the indicators that are not part of the visions are imputed using the PCA based imputation methodology from above, again with five PCs. This set of indicators is the “most likely” outcome if the visions of the city of Vienna are reached since it uses the underlying structure of indicator data set to “guess” the value of the missing indicators based on the indicator values included in the visions of the city.

We compare this scenario with other scenarios where the missing indicators are imputed first with the current values, i.e. with the assumption that the other indicators would not change and in the second scenario with a linear prediction of each indicator using the historical time series of the indicator. The latter can be seen as a “continue as before” scenario as the indicators not included in the visions develop like they did over the last 5 decades.

The resulting visions can be seen in Figure 7. The Figure shows the historic (already available) data in black. The smaller points correspond to earlier years. The coloured points indicate possible future developments and include the visions of 2030 and 2050 respectively. The indicators that are not part of the visions are imputed in different ways: The green dots are points, where the imputation is drawn from a linear forecast of the current development of the indicators. The orange (2030) and red (2050) dots are the visions complemented by the current values for the missing indicators. Finally, the blue dots (light blue 2030, blue 2050) are the visions together with the imputation via PCA described above.

As mentioned above it is expected that these blue dots show the most likely indicator combinations based on the indicators, included in the cities’ vision, since the imputation uses the underlying structure of the indicator data. One can see that the green dots which somewhat identify the “continue as before strategy” to city development are quite far off the blue data points in particular in the second PC, indicating that some effort is needed to reach the goals in the visions (according to Figure 5 mainly in the societal and safety indicators). One can also see that the red points are not surprisingly on the path from the current data points towards the blue points since they are close to the last data points in the historical data. This direction could be considered as an input for the definition of corridors where indicators have to be directed towards the goal scenarios.

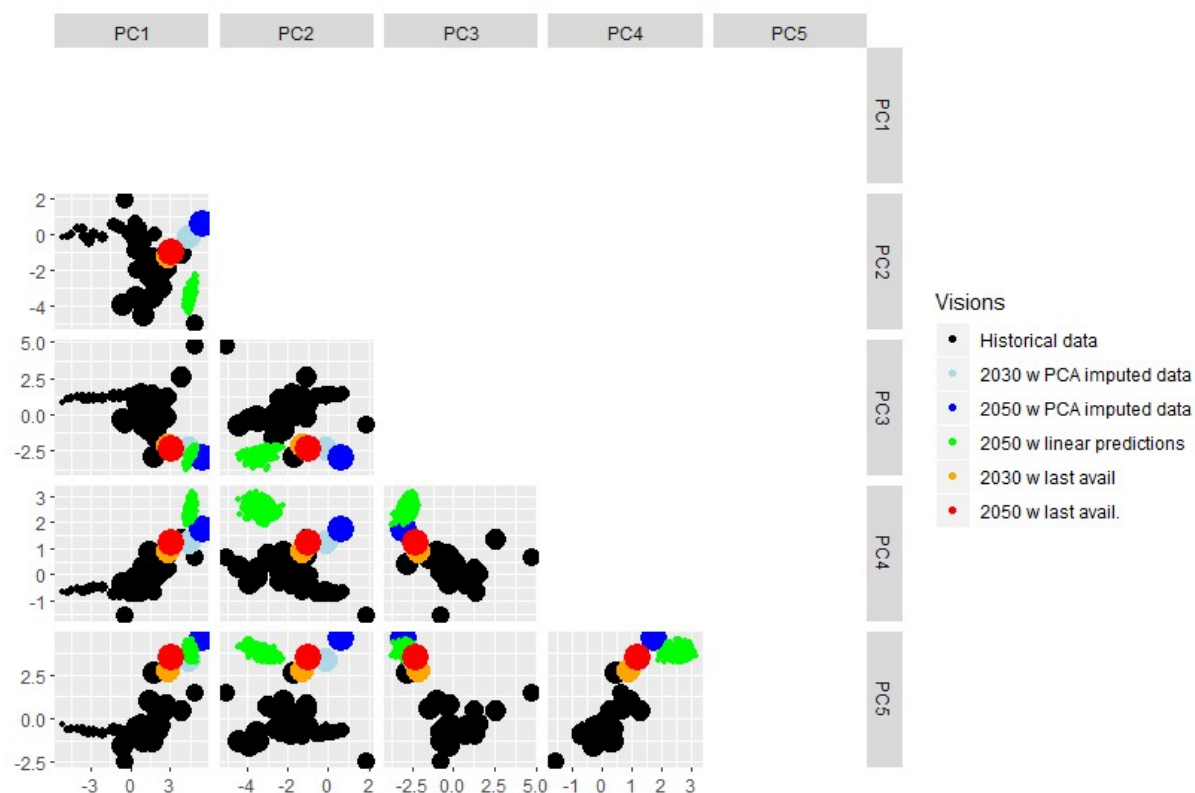


Figure 7: Visions for Vienna - First five PCs for the data of the city of Vienna. The visions together with different forms of imputed data for the indicators not included in the visions are added in different colours. The year of origin of the data is given by the size of the data points with small points coming from older data. One can again see that the in particular in the first PC the older data is more to the left of the graphs and the newer data more to the right indicating a growing value in the first component (improvement in the indicators with positive first factor loading). The most likely scenario for the visions is given in blue, the "continue as before" scenario is given in green.

## 6.2 Results of collaborative filtering

### 6.2.1 Similarity of indicators

As outlined in section 5.3, one of the attractive features of the collaborative filtering / matrix factorisation approach lies in the *geometric interpretations* in embedding space which will be shown and discussed in this section. In order to visualize this mapping of indicators as well as of geo-entities, a further dimensional reduction is required – which is typically done by PCA (similar as for the results presented in the last section).

Here a PCA is performed on the 20-dimensional embedding space, and the first and second PCA component (PCA1, PCA2) are used (for most of the figures) as axes in a 2D Plot. Note that PCA1 and PCA2 are linear combinations of all the latent factors, but they could be interpreted as latent factors themselves, contributing either positively or negatively to the indicators.

The picture below (Figure 8) shows the mapping of LEVITATE indicators in this 2D space, coloured according to their assigned LEVITATE dimension – Safety, Society, Environment, Economy <sup>4</sup>. If an indicator is in the left half of the plot (left of the y-axis) - this means that PCA1 contributes negatively to the corresponding indicator. The four ellipses (with their centres indicated by big dots) indicate the “regions” for each dimension where the majority of corresponding indicators is located. It is evident that the “economy” region is shifted left compared to the others (i.e. PCA1 contributing negatively). On the other hand, PCA2 contributes slightly positively to all four dimensions.

Indicators which are similar to each other (showing positive correlation), are expected to have a “small distance” in embedding space; their vectors in embedding space point into a similar direction (expressed by high cosine similarity - or small cosine distance <sup>5</sup>). This can be observed in the plot for example for following groups of indicators:

- Fatalities\_1,2,3
- PercSafety\_1,2,3
- BuildingVol\_1,2,3
- PopDens\_1,2
- Income\_n
- GINI\_1,2

Also, indicators within one dimension can be expected to be closer to each other than across dimensions (even if several goals within *one* dimension can also be conflicting). In fact, the average cosine distance has been calculated and compared between indicators *within* one dimension (0.86) to the average cosine distance between indicators *across* dimensions (1.0) – which is the expected result.

---

<sup>4</sup> Indicators not labelled have been included in the calculation and are assigned to a dimension and goal but have not been selected as LEVITATE indicators.

<sup>5</sup> cosine distance between two vectors is defined here as  $1 - \cos(\theta)$  where  $\theta$  is the angle between the two vectors, refer to <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.cosine.html>.

Note also that Injured seems to be negatively correlated to Fatalities, confirming the (weak) evidence discussed in the previous section. Similar for Income (Income\_n) against CO2 or against energy consumption for transport (EnergyConsumTran).

Finally, a lot of yellow dots can be observed in Figure 8. These correspond to the many distinct geo-entities for one particular year where measurements are available. This illustrates the fact that geo-entities can be represented in the same space. The structure of these geo-entities in embedding space will be analysed in more detail in the following subsection.

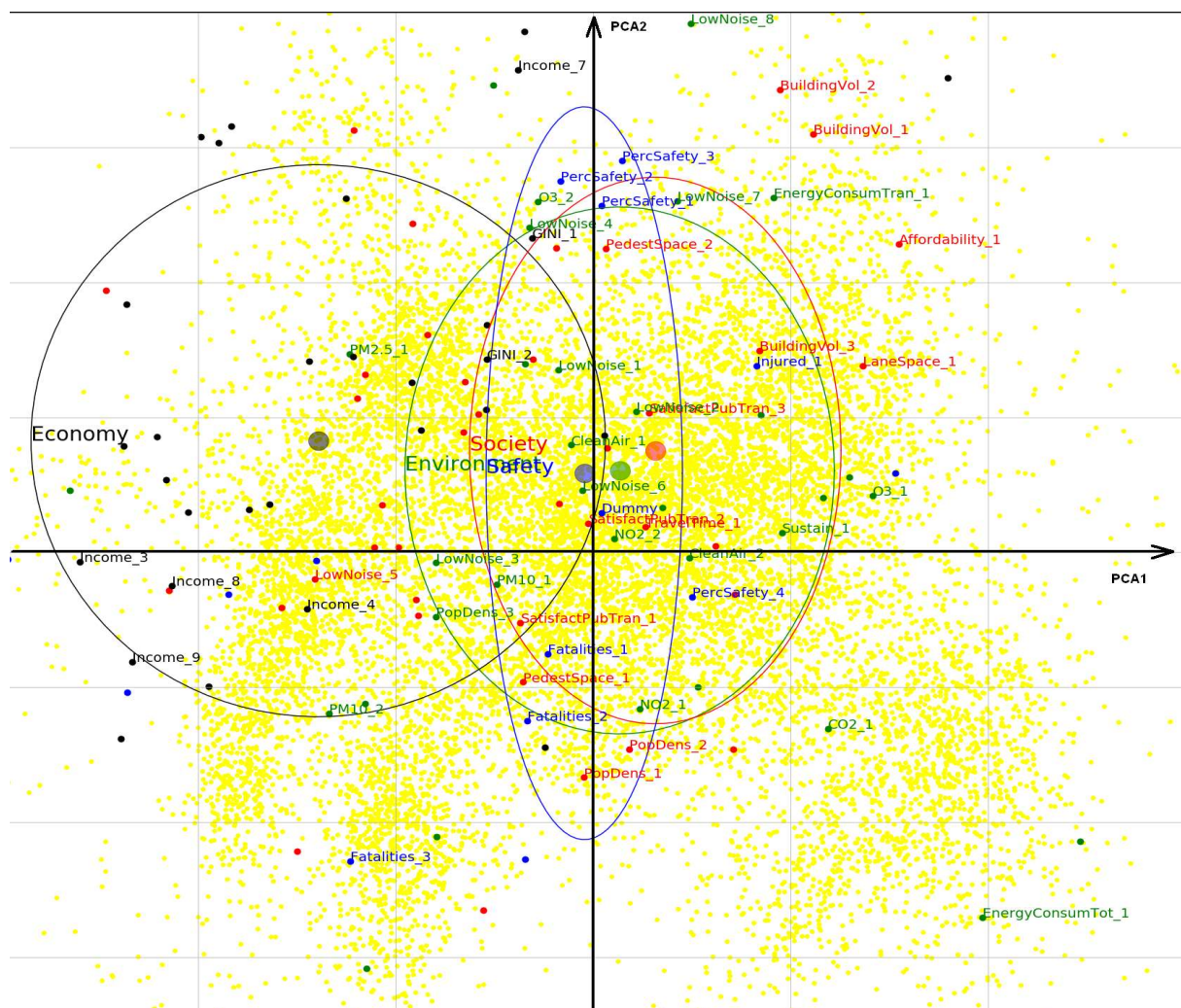


Figure 8: Visualisation of LEVITATE indicators and four LEVITATE dimensions in embedding space (first two primary components). Geo-Entities are also shown as yellow dots.

## 6.2.2 Similarity of geo-entities

As previously seen, indicators which are similar to each other, are mapped close to each other in a lower-dimensional (PCA) representation of the embedding space. The same is



true for the mapping of geo-entities (or more precisely, geo-time-entities, i.e. geo-entities in one particular year).

Figure 9 shows the similarity of geo-entities belonging to the same region, indicated by colours. For this illustration, only the data points from the year 2000 to the present have been used. The figure shows two 2D-plots: PCA1, PCA2 (i.e. the same axes as used in Figure 8) and PCA1, PCA3. The colours used are:

- Red: Central Europe
- Blue: UK
- Green: Northern Europe
- Cyan: Southern (& South-East) Europe
- Yellow: Eastern Europe
- Magenta: Central Asia

By comparing the two different projections, it is easier to identify clusters. The blue and cyan points, for example, are completely overlapping in the upper plot (PCA1, PCA2) but can be clearly distinguished in the lower one (PCA1, PCA3).

### 6.2.3 Development over time and identifying of visions

As already explained, geo-entities *move* in the embedding space as time passes by. Depending on the amount of input data (indicator values) available for a specific year, the location is subject to certain fluctuations, which can be smoothed by averaging over a longer time interval (e.g. a decade).

Finally, along with this development over time, the possible visions are analysed for Vienna, based on the target values discussed in section 4.1, in the context of the applied collaborative filtering approach.

Figure 10 illustrates several aspects in one diagram: LEVITATE Indicators are again shown in -two-dimensional projection (first two PCA components) of the embedding space, with colours indicating the dimension and confidence ellipses indicating the corresponding region in the subspace for each dimension. Further, the distribution of geo-entities is shown by yellow dots. For the example of Vienna (represented by NUTS-2 code "AT13"), the development over the decades (196x – 201x) is shown by cyan dots (the values have been obtained by averaging over all data points available in that decade). It can be observed that the movement over time in this embedding space is quite steady. Finally, the two fictive entities AT13\_2030 and AT13\_2050 (representing the Vienna visions for 2030 and 2050) are also shown in this diagram – it can be observed that these are clearly not on a continuation "curve" of the real historical values.

Comparing these results to the vision mappings presented in the last section (in Figure 7), they are similar to the "visions together with the imputation via PCA" (the blue dots), and significantly different from a "continue as before strategy".

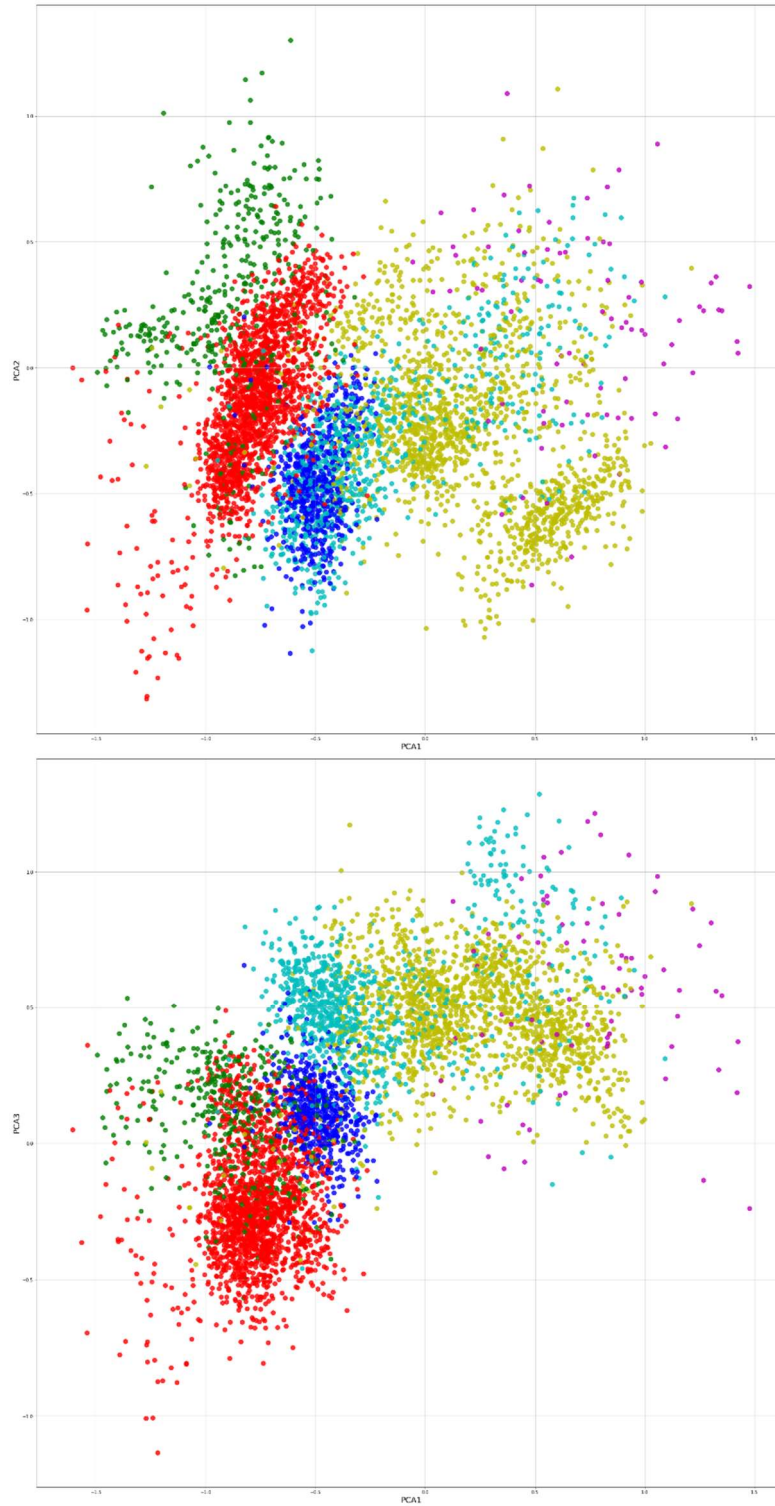
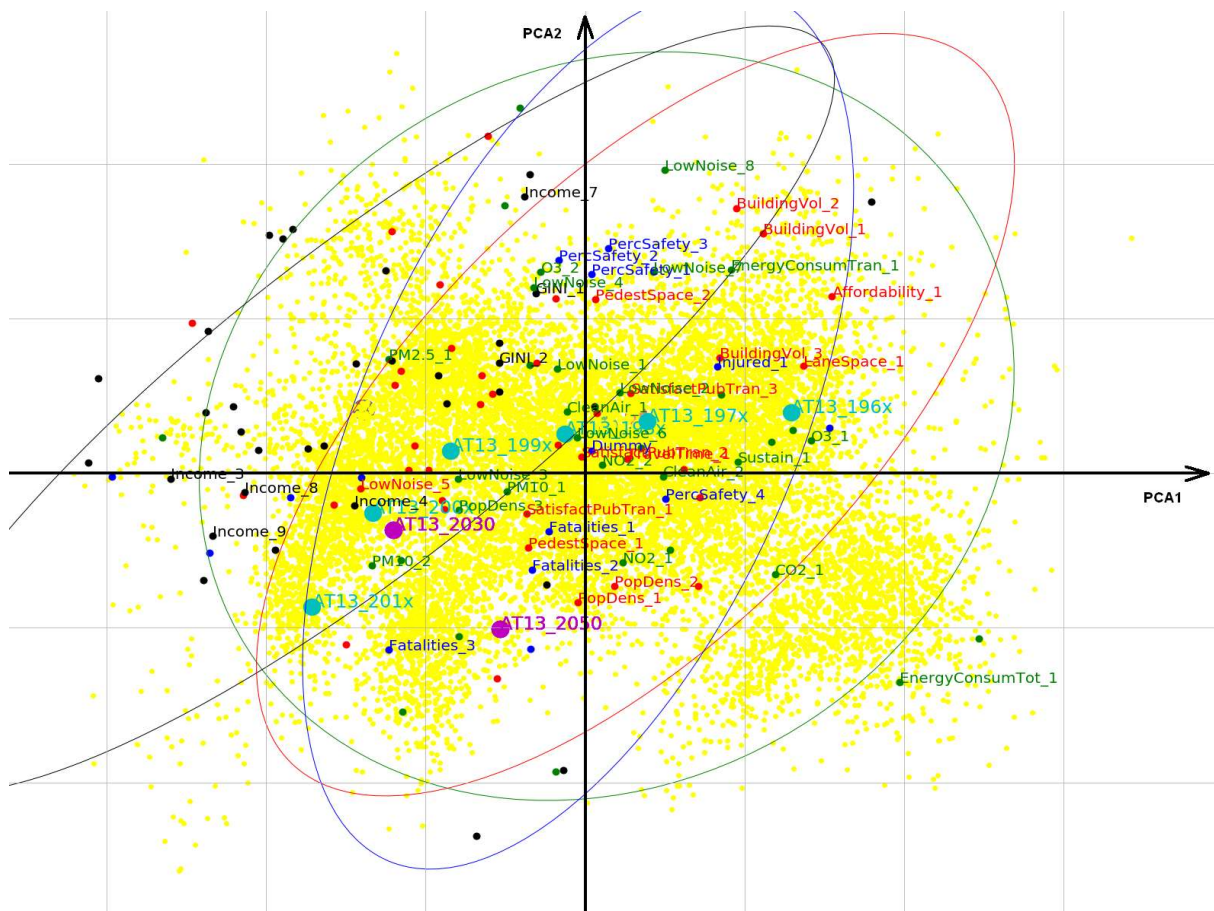


Figure 9: Illustration of clustering for geo-entities according to European regions in two-dimensional projection of the embedding space (top: PCA1, PCA2; bottom: PCA1, PCA3). Colour legend is explained in text.



### 6.3 Common interpretation of results

The results from applying two different approaches have been shown in this chapter, highlighting aspects like the similarity of indicators, similarity of geographic entities, development over time and the identification and visualisation of visions.

As a general observation it can be stated that investigating these relationships in a quantitative way, based on the scarce data available, even more as they are collected on different geographical levels, represents a big challenge. Some promising results have been obtained so far, but further cross-checks are definitely needed to support the further work in LEVITATE based on these results, mainly the transformation paths towards desirable visions and the backcasting process.

Comparing the two approaches, it is found that

- Similarities between indicators (i.e. strong correlation, but also anti-correlation) could be identified by both approaches.
- Clustering of geo-entities is quite strong and can be demonstrated in both approaches – geo-entities of same region are close to each other in parameter or embedding space.

- Developments over time (how geo-entities change over the decades) are also clearly visible.
- There are several ways how to map and illustrate a concrete vision, with specific target values for a city or region, with slightly different but consistent results.

There are also clear limitations of both approaches that should be mentioned here:

- Any visualisation shown here is based on dimensional reduction by PCA and shows a two-dimensional projection only. This can lead to mis-interpretation in some cases. (For the collaborative filtering approach, the “full” embedding space can still be considered for any quantitative calculations, e.g. regarding the distance between current representation of a city and its “vision”, and the recommended direction towards this vision.)
- The PCA based data imputation in the statistical analysis and the matrix factorisation in the collaborative filtering approach both have an uncertain impact on the results shown here. Any quantitative outputs should therefore be seen with great caution.

# 7 Conclusions and outlook

## 7.1 Identifying visions in City strategies

This deliverable has described

- the process of collecting open data corresponding to LEVITATE indicators,
- defining visions based on these indicators,
- statistical and machine learning approaches to exploit these data for correlations and patterns, and finally
- the results of data analysis, fostering better understanding of the high-dimensional indicator space – how certain geo-entities are located and moving in this space.

The most important part for the backcasting approach in LEVITATE is the linking of these results to the (already specified) *visions* of cities and regions – as these visions represent the starting points for backcasting.

Table 12 summarises the mapping of LEVITATE goals and indicators to key quantitative targets that can be used to identify a vision in LEVITATE context, for the two examples of Vienna and Greater Manchester.

Defining a quantified vision by a (prioritized) set of goals and targets in a formal way seems to be straight forward. It is clear, however, that in reality this might be a quite lengthy and complex process. With the approaches followed in this deliverable, it has been demonstrated that it is possible to identify “regions” in indicator space that are close to such an idealized vision and consistent in terms of correlations between various target indicators – despite the limitations which are due to the high sparsity in the available data set.

## 7.2 Identification of feasible transformation paths

Once these visions are defined and described, the next step is the specification of possible transformation paths. This process connects the results of this deliverable (and the outputs of tasks T4.1 and T4.2) to the preliminary results of other work packages (WP3, WP5-7, WP8) and further involvement of the stakeholders.

The core activity in WP4, based on the data analysis presented in this deliverable, will be data driven support for the specification of these paths. The observed development over time during the recent past indicates the current “velocity vector” in parameter or embedding space for a certain city or region. This can be compared to the “direction” towards the desired vision – the difference of these two vectors can be interpreted as required “change of direction”. For one or a few indicators to be optimized, this gives a simple and intuitive picture; in a high-dimensional space of correlated (depending) indicators, however, any visualisation is difficult and might lead to mis-interpretation.

The next step is the mapping to the necessary CATS parameters that are connected to indicators via impact relationships: For example, an increase in “protection of human life”

Table 12: Mapping of LEVITATE goals and indicators to quantitative targets defining a vision

Dimension	Policy Goal	Indicator	Target Vienna	Target Greater Manchester
Safety	Protection of Human Life	Number of injured per million inhabitants (per year)	(decline)	as close as possible to zero (2040)
		Number of fatalities per million inhabitants (per year)	(decline)	as close as possible to zero (2040)
Society	Use of public space	Lane space per person		
		Pedestrian/cycling space per person	(increase)	
Environment	Clean air	Emissions directly measurable: SO <sub>2</sub> , PM <sub>2,5</sub> , PM <sub>10</sub> , NO <sub>2</sub> , NO, NO <sub>x</sub> , CO, O <sub>3</sub>	Greenhouse gas emissions -50% (2030), -85% (2050)	Robust low carbon pathway to 2050 at which Greater Manchester can become carbon neutral.
	Sustainable behaviour	Rate of energy consumption per person (total)	-30% (2030), -50% (2050)	
		Rate of energy consumption per person (transport related)	-40% (2030), -70% (2050)	Sustainable modes (walking, cycling or public transport) will increase from 39% in 2019 to 50% in 2040
Economy	Prosperity	Taxable income in relation to purchasing power	(increase)	
	Fair distribution	GINI index	(decline)	

(expressed by decrease in Fatalities, Injured) could be connected to a rise of market penetration level of SAE Level-5 according to an established dose-response curve.

Finally, as the last step in this chain, a mapping to possible sequences of policy interventions will be performed that influence these CATS parameters in such a way that a desirable vision can be reached, completing the formal backcasting relationship.

## 7.3 Backcasting process

According to current state of analysis and alignment with other work packages, the actual backcasting in LEVITATE will take place on two different levels:

1. *Static component* - Case studies for cities (this will be the main subject of task T4.3) based on the following inputs:
  - a. City vision (as already documented for two examples in this deliverable)
  - b. Historical data & data driven modelling (support for defining feasible corridors)
  - c. (Iterative) Dialogue with stakeholders in order to define feasible paths of intervention towards the vision

The output of this task can be used for further investigations and verification (e.g. simulations) in WP5-7 (to be specified more precisely in task T4.4); the final result will be a Case study report including also generalized findings and recommendations.

2. *Interactive component* - within the PST Estimator, based on user interaction:
  - a. Select a (customisable) vision template
  - b. Select a (customisable) city template
  - c. If the vision cannot be reached (i.e. if the PST forecast *without* additional interventions results in a state that is not within the calculated corridor towards the vision), the user selects a set of interventions
    - i. Manually (from a full list) or
    - ii. Automatically (PST sets the interventions) or
    - iii. Semi-automatically (PST generates a short list to choose from)

These two approaches for backcasting are not completely independent from each other. In particular, it is expected that core findings of the static backcasting process (tasks T4.3 and T4.4), like generalized recommended policy interventions, will be taken over as PST features for the interactive backcasting support.



# References

- Asif, M. T. e. a., 2013. *Low-dimensional models for missing data imputation in road networks*. s.l., s.n.
- Brunner, S., 2016. A backcasting approach for matching regional ecosystem services. *Environmental Modelling & Software* 75, pp. 439-458.
- Cheng Guo, F. B., 2016. *Entity embeddings of categorical variables*. [Online] Available at: <https://arxiv.org/abs/1604.06737>
- Danciulescu, V. et al., 2015. Correlations between noise level and pollutants concentration in order to assess the level of air pollution induced by heavy traffic. *Journal of Environmental Protection and Ecology* 16, No 3, pp. 815-823.
- EC, 2019. *COMMISSION STAFF WORKING DOCUMENT: EU Road Safety Policy Framework 2021-2030 - Next steps towards "Vision Zero"*, s.l.: s.n.
- Elvik, R. etc., 2019. *A taxonomy of potential impacts of connected and automated vehicles at different levels of implementation. Deliverable D3.1 of the H2020 project LEVITATE.*, s.l.: s.n.
- Elvik, R., 2000. How much do road accidents cost the national economy?. *Accident Analysis and Prevention*, p. 849–851.
- Eurostat, E. ./., 2019. *Eurostat*. [Online] Available at: <https://ec.europa.eu/eurostat/about/overview>
- fast.ai, 2019. *fast.ai - Making neural nets uncool again*. [Online] Available at: [fast.ai](https://fast.ai)
- GMCA, 2019. *The Greater Manchester Strategy Outcomes Framework*. [Online] Available at: <https://www.gmtableau.nhs.uk/t/GMCA/views/GMSLandingPage-October19/GMSLandingPage?.linktarget=self&:isGuestRedirectFromVizportal=y&:embed=y>
- Herlocker, J., Konstan, J., Borchers, A. & Riedl, J., 1999. *An algorithmic framework for performing collaborative filtering*. s.l., s.n.
- Hinton, G. E. a. R. R. S., 2006. Reducing the dimensionality of data with neural networks.. *Science*, pp. 504-507.
- Ingvardson, J. B. & Nielsen, O. A., 2019. The relationship between norms, satisfaction and public transport use: A comparison across six European cities using structural equation modelling. *Transportation Research Part A: Policy and Practice*, pp. 37-57.
- Jacob, D., 1999. *Introduction to atmospheric chemistry*. s.l.:Princeton University Press.



Josse, J. a. F. H., 2012. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2), pp. 79-99.

Koren, Y., Bell, R. & Volinsky, C., 2009. Matrix Factorization Techniques for Recommender Systems. *Computer Vol. 42 Issue 8*, pp. 30 - 37.

M. A. Saif, M. M. Z. A. T., 2018. *Public Transport Accessibility: A Literature Review*, s.l.: s.n.

Manchester, G., 2019. *5-Year Environment Plan for Greater Manchester*, s.l.: s.n.

Mohan, D., Bangdiwala, S. I. & Villaveces, A., 2017. Urban street structure and traffic safety. *Journal of Safety Research*, pp. 63-71.

Najaf, P., Thill, J.-C., Zhang, W. & Fields, M. G., 2018. City-level urban form and traffic safety: A structural equation modeling analysis of direct and indirect effects. *Journal of Transport Geography*, pp. 257-270.

OECD, 2019. *Organisation for Economic Co-operation and Development (OECD)*. [Online] Available at: <http://www.oecd.org/about/>

Pearson, K., 1901. *Philosophical Magazine*. 2. *On Lines and Planes of Closest Fit to Systems of Points in Space*, pp. 559-57.

PyTorch, 2019. <https://pytorch.org/>. [Online].

Rode, P. et al., 2014. Accessibility in Cities: Transport and Urban Form. *NCE Cities Paper 03, LSE Cities, London School of Economics and Political Science*.

SAFiP, 2019. *SAFiP - Systemszenarien Automatisiertes Fahren in der Personenmobilität*. [Online]

Available at: <https://projekte.ffg.at/projekt/2929372>

Schölkopf B., S. A. M. K., 1997. Kernel principal component analysis. In: *Lecture Notes in Computer Science*, vol 1327. Heidelberg: Springer.

Soria-Lara, J. A., 2017. Participatory visioning in transport backcasting studies: Methodological lessons from Andalusia (Spain). *Journal of Transport Geography* 58, pp. 113-126.

TfGM, 2019. *Greater Manchester Transport Strategy 2040*, s.l.: s.n.

UN, kein Datum *SDG Indicators Database*. [Online]

Available at: <https://unstats.un.org/sdgs/indicators/database/>

Vergragt, P. J., 2011. Backcasting for sustainability: Introduction to the special issue. *Technological Forecasting & Social Change* 78, p. 747-755.

Vienna, C. o., 2015. *Urban Mobility Plan Vienna*. [Online]

Available at: <https://www.wien.gv.at/stadtentwicklung/studien/pdf/b008443.pdf>

WDI, 2019. *The world bank*. [Online]

Available at: <https://www.worldbank.org/en/about>

Wien, M. d. S., 2019. *SMART CITY WIEN - Rahmenstrategie 2019 - 2050*, s.l.: s.n.

Zach, M. e. a., 2019. *Definition of quantified Policy Goals. Deliverable D4.1 of the H2020 project LEVITATE.*, s.l.: s.n.

Zahabi, S. A. H. et al., 2012. *Transportation Greenhouse Gas Emissions and its Relationship with Urban Form, Transit Accessibility and Emerging Green Technologies: A Montreal case study*, s.l.: European Working Group on Transportation.

Zanden, J. L. v. et al., 2014. *How Was Life?: Global well-being since 1820*, s.l.: OECD Publishing.

# Appendix

## Used Terminology

Following definitions that have been discussed in LEVITATE across the work packages are relevant for this deliverable; these are the terms that are proposed to be used throughout the project:

<b>Term</b>	<b>Description</b>	<b>Examples</b>
Impact categorization	In order to simplify the categorization of CATS impacts, two main categories are identified:  (1) Direct impacts: impacts that are produced directly from the introduction of CATS on the transport system such as vehicle design and driving behaviour.  (2) Indirect impacts: impacts that are a by-product of the direct impacts of CATS. For example, driving behaviour will affect road user interaction and therefore road safety which is an indirect impact.	
Policy	Definition: A set of ideas or a plan of what to do in the future in particular situations that has been agreed to officially by a group of people, a business organization, a government or a political party.	Environmentally friendly, social equity, increase in health, liveability
Policy goals / Policy objectives	Definition: A single target within the whole policy (should be SMART)  Should be third order impacts, which are wider impacts e.g. societal and are usually not directly transport related.	One of the European 20-20-20 Targets:  The 2020 energy goals are to have a 20% (or even 30%) reduction in CO2 emissions compared to 1990 levels.
Policy interventions / measures	Definition: An intervention is an action undertaken by a policy-maker to achieve a desired objective. Interventions may include educational programs, new or stronger regulations, technology and infrastructure improvements, a promotion campaign.	Introduction of a city toll, conversion of driver license training, dedicated lanes for automated vehicles
Vision	Definition: Description of a future situation defined by a bundle of vision characteristics and dedicated at a specific point in time.	The case of Vienna (modal share, mobility demand, penetration rate of

	Note that this term is used instead of the term “desired future scenario” that was used in the project proposal, in order to avoid any confusions with simulation scenarios in LEVITATE context	automated vehicles of level x, ...)
Vision characteristic	Definition: An indicator representing a policy goal that has to be achieved at a certain time. A single target within the vision in the level of first and second order impacts (which occur in the transport system, on a trip-by-trip basis / which involve system-wide changes in the transport system).	Penetration rate of automated vehicles of level x, population density, number of near miss / collisions, Number of accidental deaths, particulate pollution, noise, public green space.
Transformation Path	Definition: A postulated sequence or development of policy interventions / measures (and external events/measures/conditions) driving from a vision 'A' at time 'X' (which can be the current situation) to a vision 'B' at time Y.	Situation now in Vienna (modal share, mobility demand, penetration rate of automated vehicles of level x, ...), measures: campaign in 2020, funding for dedicated research in 2025, restricted access to freight in 2025, city toll in 2028; situation in 2030: (specified modal shift, expected mobility demand, penetration rate of automated vehicles of level x, ...)