Mobility of the Future: Typologizing Global Cities for the Simulation of Future Urban Mobility Patterns and Energy Scenarios

by

Sean Hua

Submitted to the Dept. of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Author .....
Dept. of Electrical Engineering and Computer Science
August 11, 2017

Certified by .....
Moshe Ben-Akiva
Edmund K. Turner Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Christopher Terman
Chairman, Masters of Engineering Thesis Committee

Mobility of the Future: Typologizing Global Cities for the Simulation of Future Urban Mobility Patterns and Energy Scenarios

by

Sean Hua

Submitted to the Dept. of Electrical Engineering and Computer Science
on August 11, 2017, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The MITEI-sponsored *Mobility of the Future* project sets out to create a viable framework for analyses and predictions of urban transportation behavior in response to inevitable changes such as improved vehicle technologies, emergence of novel transit services, and policy changes motivated by population growth and emission control. In order to feasibly simulate these scenarios on a global scale, we need to first determine a few prototypical cities that best represent the entire world, each exhibiting qualities that encompass the group to which it belongs. Our methodology for accomplishing this is centered around *machine learning*. After collecting and pruning relevant, up-to-date data, we perform dimension reduction and clustering to ultimately generate appropriate prototype cities. These cities will be used as test beds for future mobility scenario exploration and analyses.

Thesis Supervisor: Moshe Ben-Akiva
Title: Edmund K. Turner Professor of Civil and Environmental Engineering

Acknowledgments

First and foremost, I would like to thank Jimi Oke, without whom this would not be possible. Your relentless work ethic and contagious passion have truly inspired me over the past year. To Carlos, I am grateful for your honest feedback and constructive guidance throughout this journey. To Professor Moshe Ben-Akiva, thank you for giving me this opportunity. Shout out to Yafei and Joey for your key contributions, and to Michael, Akshay, and the rest of the ITS Lab for good vibes, edifying seminars, and free lunches!

Cheers to Sul Gi, who has helped me through the ups and downs. To my parents, thank you so much for your unconditional support throughout all these years - I would not be here without you. I love you both so much, and this thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Having spent four years of my undergraduate career working on projects (in class as well as industrial internships) strictly in the realm of computer science, I wanted to stray from this path a bit as I set out to complete my thesis for a Masters degree and apply all that I have learned to an endeavor in another field. I was curious about how I could contribute to a project whose motivations were markedly different from those that I was familiar with, and eager to learn from a discipline largely novel to me.

Mobility of the Future is a project deeply rooted in transportation and energy, both of which are seemingly unrelated to computer science. However, in this day and age, computer science has become so versatile and widespread that it is an essential tool in a variety of educational fields, and transportation and energy research is no exception. Fast-forward one year, and here I am writing about all that I have learned about urban mobility, witnessing firsthand how computer science concepts can be seamlessly applied to achieve insightful results in a vastly different context.

1.1 Project Background

Mobility of the Future is a project affiliated with the Intelligent Transportation Systems Lab at MIT and is part of a bigger operation under the directive of the MIT Energy Iniative. We are sponsored by numerous companies in the transportation and

energy sector [31], all of whom are concerned about the ever-changing landscape of urban mobility, especially in today's world of rapid technological advancement and heightened awareness of the implications population growth and climate change can have on transportation policies.

Our goal is to develop a framework that can accurately assess mobility behavior in response to the aforementioned changes for cities around the world, providing a versatile and data-driven tool that the companies can use to gain better insight into how the future will impact their businesses. However, there is much more value in this project than just the end product, considering the vast amounts of data collection and analysis that will have to be done to reach that point. In fact, one of the most valuable results stemming from our efforts is the carefully curated cities dataset.

1.2 Development Plan

To holistically capture the global scope requires documenting every single city in the world and creating a separate model for each. However, this is not very realistic and is clearly not the most cost-effective approach towards our goal. Instead, what we concluded was an alternate viable plan is to come up with a few prototype cities that best represent groups of similar cities across the earth, each with its own unique characteristics in economy, mobility, and environment/health. With a justified collection of archetypal cities available, we can adapt and calibrate a model for each of them and draw conclusions for every city based on which cluster they were ultimately placed in.

In order to obtain these clusters and their corresponding classifications, we must first diligently collect relevant data on the cities that we want, then utilize machine learning principles and conduct numerous statistical analyses on the results for verification. There are countless design decisions that must be made before we dive into execution, and they will be explained in more detail over the next few chapters.

1.3 Previous Work

There are a few studies that have attempted urban typologization, and although they do contain potentially valuable information, there is still a lot of room for improvement as we strive to create our own typologies that are tailor-made for this project. Some of the past studies are based on outdated information, others are focused on fields outside of mobility, and a couple even utilize questionable methodologies resulting in unreliable results. Consequently, we decided to design and execute our own pipeline to create typologies that are not only robust but also up-to-date with the latest data. Nevertheless, here is a selection of past published work that we found interesting and potentially useful.

1.3.1 Harris

Remarkably, in 1943, Chauncy Harris conducted a study and published *A Functional Classification of Cities in the United States* [27]. Harris, a visionary urban geographer, recognized that literature during his time failed to truly classify cities correctly, so he set out to conduct his own. Collecting data mostly pertaining to employment and occupation within the United States, he attempted to categorize each city as manufacturing, retailing, diversified, wholesaling, transportation, mining, educational, resort/retirement, or others. Although Harris' effort was admirable in classifying nearly 1000 cities, the lack of a modern scientific approach, the obsolete nature of the data, and the limited scope to within the US make this an antiquated resource.

1.3.2 Louf and Barthelemy

A much more recent attempt to classify cities was done by Louf and Barthelemy, who published *A Typology of Street Patterns* [29] in 2014. However, their basis for classification was solely street patterns, yielding a one-dimensional model with results that were limited in context. They used a quantitative approach, focusing on the "probability distribution of shape factor of blocks with a given area and define what could constitute the 'fingerprint' of a city".

Ultimately, although their methodologies are scientifically sound (they performed hierarchical clustering and analyzed the subsequent dendogram), they did not justify their decision to have 4 clusters as well as the fact that one of the clusters accounted for 78% of the total 131 cities.

1.3.3 Priester

Perhaps the most similar to what we are trying to do, *The Diversity of Megacities Worldwide: Challenges for the Future of Mobility* [36] by Priester et al. attempted to analyze 41 major urban centers around the world based on 59 indicators, all related to mobility. After performing dimensionality reduction and obtaining 13 distinct principal factors, they were able to cluster the cities into seven typologies: paratransit, auto, non-motorized, hybrid, traffic-saturated, transit, and Manila (a singleton).

Despite an excellent execution plan from start to finish, where *Priester* is lacking is the number of cities included in the study as well as certain factors that we think are important such as health and environment. The study also utilizes data from 1995 (which they defend is still relevant today), but we would like to incorporate information that is much more recent. With that said, their methodologies are very sound, providing a valuable resource for us.

Chapter 2

City Selection

Before collecting data, we have to first determine which cities are of interest to us, thereby establishing our *pool of cities*. There are many different ways we can approach this, but the goal in mind should be to find a comfortable balance between maximizing global coverage and minimizing overhead and extraneous entries.

2.1 What is a City?

There are numerous interpretations of what a city is, and it is completely up to us to choose one as the basis of our exploration. Wikipedia [12] lists "three basic concepts used to define urban areas and populations" as follows:

- **city proper:** the single political jurisdiction which contains the historical city centre
- **urban agglomeration:** an extended city or town area comprising the built-up area of a central place (usually a municipality) and any suburbs linked by continuous urban area
- **metropolitan area:** one or more urban areas, as well as satellite cities, towns and intervening rural areas that are socio-economically tied to the urban core, typically measured by commuting patterns

Of course the aforementioned terms and their respective definitions are not completely rigid; different sources or studies may have inconsistent views regarding exactly what constitutes each *city* variant. For the sake of clarity in this project, we will define the three terms exactly as stated above and make a judgment on which one best suits our goals.

To recap, we want to cluster cities around the world mostly based on factors that are related to transportation and energy. The following chapter will explain exactly what data is relevant to this project, but keeping in mind the ultimate goal of simulating future transportation scenarios, encapsulating an area's entire commute system and travel patterns is paramount to providing an accurate basis for our models. Looking at the three different interpretations of what a city is, it is clear that *metropolitan area* provides the best definition.

2.2 Which Cities to Include?

Now that we have established a standard for what a city is, the next step is to determine which ones to include in our pool of cities. Ideally, we want to collect data on *all* of them so that we can operate on a legitimately holistic global scope. However, due to limited resources and time, as well as the rapidly diminishing returns as the number of cities increases, a more stringent selection is much more practical. By excluding cities that fall below a threshold for global impact or importance, we can cut the number of cities to a more manageable figure at the cost of a relatively small amount of true worldwide coverage. The problem is, this line we have to draw is quite subjective and rather abstract. In an attempt to realize it, we started with the most obvious filter — population.

By only including cities with a population of over $750,000$, we were able to reduce the number of cities down to just over 600. While this is a reasonable scale, we would still prefer to cut it down further. A quick perusal of the list of cities revealed

that some countries had huge representation, with many of the cities being either irrelevant in a global transportation/energy context or very similar to others in the same nation. After cherry picking these cities for removal, the pool of cities dropped to around 400 at barely any cost of coverage.

At this point, it was difficult to find other methods of filtration without the information of how much relevant data we have for each city, so we had to wait until the end of the data collection phase, which is the topic of the next chapter, to make further cuts. Ideally, if each city had complete data for all the indicators we were aiming for, then we would be more than happy to keep the current number. However, we ultimately observed that dozens of cities barely had any information available, so they could not be kept for analysis. It was unfortunate that many cities had to be removed due to lack of data, but we took solace in the correlation between availability of data and global prominence, which makes intuitive sense. Ultimately, we were able to narrow down our pool of cities to a final size of 330. A list of these city-country pairs can be found in Table A.1.

Chapter 3

Data Collection

Collecting useful data on the cities is arguably the toughest and most crucial part of the entire project. Because the ultimate prototype cities are generated directly by performing clustering and subsequent analysis on our dataset, how successful we are depends heavily on the quality of information that is obtained.

3.1 What Type of Sources to Consult?

Our mission from the very beginning was not only to create a set of models for assessing future global mobility, but also to compile an up-to-date and comprehensive dataset related to transportation and energy and make it available for others to play around and analyze. As a result, we decided to only use sources that were free and publicly available, which admittedly made the task more difficult. Beyond that, however, there were no further restrictions, so material ranging from professional studies to public crowdsourcing were all considered as long as there existed an assurance of reliability.

3.2 What Kind of Data to Collect?

The focus of the data collection necessarily fixated on information relevant to transportation, economy, and environment/health, but in general, we included whatever we could find because pruning could always be done upon completion of this phase.

There was also a heavy emphasis on the data being as current as possible for obvious reasons, and we rejected stale (more than five years old) data unless it was essential.

3.3 Dataset Construction

In order to actually compile a dataset filled with the cities and their indicators, I created a very basic CSV file with the city names and the country where it is located — call this *cities.csv*. When adding information to this file, we would have to take a look at each individual data source, pick out the specific fields that we wanted to append, and attach them to the corresponding city. Fortunately, most of the sources contained both city and country names and could easily be converted into a CSV file, so programmatically grabbing the information in a computer script was simple enough.

One challenge we faced, however, was that consistency of city and country names across sources was uncommon, as they would often be spelled differently. For example, some sources would use the official name of countries (*e.g. Republic of Korea*) whereas others would use *South Korea* or even *S. Korea*. Needless to say, this caused much frustration because the only fool-proof way to overcome this obstacle was to manually change the egregious city or country names in every data source to the standard ones in *cities.csv*. Additionally, the script was not one-size fit all, as it still required us to specify certain parameters in the code. Shown below is the pseudocode for the matching algorithm. My actual code took advantage of the *csv* library in Python, which allowed me to read rows from CSV files as well as write rows to a new file. In hindsight, it probably would have been easier to use the very powerful Pandas toolkit to manipulate the dataset, but the script was extremely simple to write and worked perfectly, so I stuck with it.

```
output.csv      <- create new output csv file
my_rows     <- read rows from cities.csv
new_rows    <- read rows from datasource.csv
my_headers  <- get list of indicator names from my_rows
new_headers <- compile list of indicators we want to add

append my_headers + new_headers as a new row to output.csv

for row in my_rows[1:]: # exclude top row i.e. headers
    city    <- get city name from row
    country <- get country name from row
    exists  <- False

    for row2 in new_rows[1:]:
        city2    <- get city name from row2
        country2 <- get country name from row2
        data     <- get data corresponding to new_headers

        if city == city2 and country == country2:
            append row + data to output.csv
            exists     <- True # city-country pair matched
            break out of loop due to match

    if not exists: # fill the columns with empty values
        append row + len(new_headers)*[''] to output.csv
```

There are a few key but potentially confusing points I want to go over regarding the above code. The contents of a new data source are not always entirely useful, which is why it is necessary for us to pick and choose the information that we want. This customization step is crucial and must be performed for each and every source. Often

times, a data source will not have information for a particular city in our dataset, which is represented by the boolean *exists*. If the city is there, then we will proceed as normal and append the new data; otherwise, we still want to populate the columns to maintain shape integrity of the CSV table, so we simply fill in the appropriate number of blank cells.

3.3.1 Special Sources

Most of the sources had pre-formatted data that could easily be converted into a CSV file, but a couple, namely *Degree Days (Weather Underground)* [3] and *OpenStreetMap* [18] required us to do further digging in order to obtain the data. For the former, we took advantage of their internal API to automatically download weather data, after which we compiled all the information into one file. For the latter, we found a really cool open-source Python library called *OSMnx* [21] specifically dedicated to "retrieve, construct, analyze, and visualize street networks from OpenStreetMap". This tool allowed us to generate appropriate city street networks and perform subsequent analysis to yield many useful statistics, helping us understand more about the size and density of the city's road structure, as well as connectivity of streets and frequency of intersections. Specifics regarding key statistics collected are in Table B.1 under the *OpenStreetMap* source.

3.3.2 Final State of Dataset

Ultimately, we were able to garner nearly 100 unique indicators, but obviously not every city had complete data. In fact, very few were complete, and some even had significant holes, which was to be expected when working with different sources, each with a different subset of cities represented. In the next chapter, I will talk about how we resolved this issue while prepping the dataset for clustering. For a summary of the sources and their indicators, please refer to Table 3.1. A more detailed list is available in Table B.1.

Table 3.1: Summary of data sources, indicators, years and number of cities

Source(s)	**Indicators**	**Years**	**No. cities**
Demographia [13]	population, land area, population density	2016	330
Degree Days.net [3]	heating/cooling degree days	2012-16	317
Global Bus Rapid Transit [4]	fleet size, fare, stations, system length, ridership	2010-17	301-330
Global City Indicators [16]	GDP, poverty rate, infant mortality, life expectancy	2013	93-330
Global Petrol Prices [14]	gasoline price	2017	330
Innovation Cities Index [11]	innovation score	2015	238
Internet World Stats [15]	internet penetration, digital access	2017	323-329
Numbeo [17]	indicies: cost of living, rent, groceries, purchasing power, affordability, crime, safety, healthcare, pollution, traffic (time), inefficiency, emissions, quality of life, climate; price-to-income ratio	2016	126-223
Open Street Map [18]	average (weighted) neighbor degree, clustering coefficient, circuity average, degree average, edge length, edges, intersections, nodes, street length, segments, self-loop proportion	2017	243-259
Pew Research Center [35]	smartphone penetration	2017	218
Simple Maps [19]	latitude, longitude	2017	330
Tom Tom [10]	congestion level (overall, morning peak, evening peak, highways, non-highways)	2016	146-154
Union Internationales des Transports Publics [7]	energy consumption	2015	72
United Nations Habitat [8]	Gini coefficient, CO2 emissions, air pollution density, unemployment, urbanization level	2011-14	129-330
World Health Organization [9]	road traffic deaths	2013	313
*Various	elevation, metro stations/ridership/system length, motor vehicles, modeshares	2010-17	165-330

Chapter 4

Preparation for Clustering

Our complete dataset of 330 cities with nearly 100 unique indicators is not yet ready for clustering. One issue with the current state is the presence of holes, which will inevitably lower the accuracy of clusters. To resolve this, we tried to manually look for each piece of missing information, and when that did not work, used country data as a reasonable substitute. Another concern was the large dimensionality of the dataset. Performing matrix operations on something of this size is unnecessarily resource-intensive and time-sensitive. As a result, we performed factor analysis to reduce the number of dimensions while minimizing data loss.

4.1 Pandas Toolkit

One thing that I want to first mention is Pandas, which I briefly touched upon in an earlier chapter. It is one of the most popular and powerful data analysis libraries for Python. Converting our CSV file into a Pandas dataframe is essential to taking advantage of all the available tools in the package. With this in place, we can easily manipulate our rows and columns to fit the exact structure we desire. We are also able to perform statistical analyses on the data to give us better insight into the numbers, which can prove valuable down the stretch.

4.2 Handling Holes in Data

One issue that is inevitable in any project involving data collection is incomplete information. While we did our best to brute force our way through by searching everywhere for missing data, the reality is that some information is not publicly available or simply does not exist. In cases like this, we decided to infer a reasonable estimate by using country data, which was significantly easier to find. We were wary of this approach because it could potentially encourage cities from the same country to be clustered together, but decided that it was fine for indicators that were likely to be uniform across a country, like gas prices.

Finally, as a dreaded last resort in order to complete the dataset, we had to fill the blanks with column averages, which could really skew the cluster results if there was signficant overlap of the same holes across many cities. Luckily, this was not the case, as the cities with incomplete data generally had holes in different indicators.

4.3 Removing Unusable Data

Because all of our subsequent analysis required numerical data, qualitative indicators such as *Metro System Name* or *Primary Export* were simply omitted. In addition, there were some indicators that were not relevant and had to be removed as well, like *Number of Museums*. Ensuring that the dataset is free of extraneous data is important because all the indicators are taken into account during the dimensionality reduction phase.

4.4 Normalization and Standardization

Having extracted information from a variety of different sources, we expect the data to vary wildly with respect to their units. Some of the data are numerical counts (*e.g. population*), some are averaged values (*GDP per capita*), and a few are not even raw data (*Innovation Index*). Thus, it is important to standardize them so that there

exists a common basis for analysis.

How do we normalize the data to make them balanced? The main ideology here is to constrain the numbers to a fixed range. All of the indicators are quantifiable—it is just a matter of making sure all of them represent intensities of the same scale which we define. We decided to linearly normalize the data and transform it onto a [0,10] range, which is one of many options — [0,1] would have worked just as well, for example. By doing this, we are preventing unusually large values from creating unfair biases during analysis.

4.5 Dimensionality Reduction

There are several benefits of dimensionality reduction. Reducing dimensionality is necessary for lowering the runtime of clustering algorithms to a feasible duration, as well as improving the performance of the machine learning model by removing multi-collinearity in the data [22]. Another advantage, which may seem quite counter-intuitive, is that the transformed dataset may actually hold a better representation of the true factors that define a city, depending on the methodology that is used. Here we discuss two of the more well-known dimensionality reduction algorithms: *Principal Component Analysis* and *Factor Analysis*, both of which we experimented with.

4.5.1 Principal Component Analysis

One of the most widely accepted and used techniques for linear dimensionality reduction is *Principal Component Analysis* [40]. There are a few variants of this method, but the general idea is to construct the covariance matrix of the data and calculate its eigenvectors. The ones with the highest eigenvalues are the principal components of our model, and they become the new factors for our data. Data loss is inevitable when we reduce the dimensionality from nearly 100 to the single digits, but by retaining

the most important indicators, we are confident the new data model has not forfeited any integrity. We performed Principal Component Analysis on our dataset by taking advantage of the *scikit-learn* Python library, one of the most popular resources for machine learning. Unfortunately, the resulting factors did not agree with our subjective understanding of the indicators, so we proceeded with the alternative route of *Factor Analysis*.

4.5.2 Factor Analysis

The underlying conceptual philosophy of *Factor Analysis* [41], is quite different from that of Principal Component Analysis, as it is based on a "formal model predicting observed variables from theoretical latent factors", whereas the latter extracts linear composites from observed variables. So while Principal Component Analysis literally mixes parts of the raw indicators to concoct new ones, Factor Analysis actually tries to find true factors that explain and give meaning to the values of the raw indicators. From an ideological point of view, this is much more desirable because we want to be clustering on the defining characteristics of cities that differentiate one from another.

Initially we simply utilized the *psych* package in R to help us conduct *Exploratory Factor Analysis* on the dataset. Using principal-axis as the factoring method and an oblique rotation, the results were better than those from Principal Component Analysis. However, our methodology assumes that the dataset values per column are normally distributed, which is definitely not true for some of the indicators. Specifically those related to *metro*, *bus rapid transit*, and *bikeshare* were dominated by 0s due to many of the cities lacking those transport modes. Subsequently, we had to explore another route within factor analysis that allows us to counter these zero-skewed columns.

We decided to treat the aforementioned indicators as censored variables by fitting them with a left-censored Tobit model [33] as below:

$$\boldsymbol{y}^* = \boldsymbol{v} + \boldsymbol{\Lambda\eta} + \boldsymbol{\epsilon} \tag{4.1}$$

$$y_j = \begin{cases} y_j^* & : 0 < y_j^* \\ 0 & : y_j^* \leq 0 \end{cases} \tag{4.2}$$

$$\tag{4.3}$$

y^* is the latent variable here, while y is the raw measurement. In the case of the zero-skewed indicators, y^* will have a negative value where $y = 0$. By using this factor model to produce our latent variables, we are able to resolve the issue of zero-dominated columns.

We then conducted Exploratory Factor Analysis in a software tool called MPlus, which allowed us to censor the necessary variables. Using maximum likelihood for our factoring method as recommended by Kamakura [28], we were also able to deal with missing variables, which meant that we could leave holes in our dataset. Of course the fewer holes the better, but by foregoing the step of using column averages to complete data, we could more comfortably maintain integrity of the data. Again, using oblique rotation to encourage commonality between indicators, we can then attempt to find the best set of resultant factors.

To determine the optimal number of factors, one of the most widely used methods is the scree test [23]. By plotting the eigenvalues in descending order against the number of factors, we can make a judgment call about where to make the cutoff. Usually the point at which the values level off is where you want to stop, after which the addition of new factors does not contribute much to data coverage. In this case, that number appeared to be around seven.

At this point, we needed to analyze the factors and try to qualitatively describe

them as the true factors that define a city. By looking at their numerical scores, we can observe which raw indicators are associated with each factor and work our way from there. This is where visualizing the data can help immensely, so we created a loadings plot of the "contributions" of each indicator to each factor (Figure C-1). By looking at this chart, we can easily connect the relevant indicators to each factor and begin to describe them. One very important point is that from this graphic, we can see that the sixth factor, or $V7$ in the figure, is essentially a non-factor due to low loadings across the board. This led us to omit this factor and instead conclude with the following six true factors:

- **BRT (*V2*):** variables related to bus rapid transit (demand, size, fleet, fares); moderate climate index
- **Metro (*V3*):** urban rail/metro propensity (demand, size, age)
- **Development/Innovation/Motorization (*V4*):** car ownership; internet penetration; development indices; smartphones; emissions
- **Congestion/Population (*V5*):** traffic/congestion; population; network nodes and inefficiency
- **Bikeshare/Sustainability (*V6*):** bikeshare; low CO_2 emissions; safety; low traffic; low network degree average
- **Network Size/Density (*V8*):** high intersection/street density; low clustering coefficient

Chapter 5

Clustering

The goal of the clustering phase is to find a balance between encouraging similar cities to be grouped together and ensuring that the clusters are different enough from one another. The best approach to achieve robust and stable groups is through unsupervised learning, and we tried many variations of the classic k-means algorithm as well as hierarchical clustering. With the availability of numerous powerful open-source machine learning libraries, especially in Python and R, we had all the necessary resources to conduct countless experiments in search of the best clustering results.

5.1 Centroid-based Clustering

The family of centroid-based clustering, of which the k-means and k-medoids algorithms belong, is probably the most well-known method. By effectively emulating a nearest-neighbors classification, the clustering works by solving the NP-hard optimization problem of minimizing the squared distances between cluster centers and cluster members across all clusters [30]. While this method works well in many cases, it may not be the best for our project because it relies on a preset k number of clusters and is influenced substantially by initial seedings of the cluster centers, both of which we do not know and should not introduce blindly. Nonetheless, we still gave it a shot as it is quite easy to implement with the *scikit-learn* package in Python.

As mentioned above, one major weakness of k-means is that we do not know the optimal k. As a result, similarly to how we determined the optimal number of factors during factor analysis, we can plot some graphs and run some tests. There are a variety of different metrics we can analyze, including the *Bayesian information criterion* [34], the *gap test* [39], and the *silhouette test* [37]. Without getting into too much detail, the Bayesian information criterion and silhouette both return a value averaged over the number of clusters, so we are looking to maximize and minimize, respectively. On the other hand, the gap statistic is a measure of the absolute fit of the model, so it will expectedly be monotonically increasing as the number of clusteres gets larger. In this case, we are doing something very similar to the scree test — trying to find at which point the goodness of fit levels off. Unfortunately, consensus among all three methods was hard to come by, likely because of the inconsistency of initial seedings, so we looked for an alternative approach that would yield a deterministic result.

5.2 Hierarchical Agglomerative Clustering

The algorithm that we turned to was *hierarchical agglomerative clustering* [32], which does not require initialization or specification of the number of clusters. Additionally, it has the advantage of offering numerous different partitions on the data depending on the resolution we want, as opposed to k-means, which yields only one. However, we must make the design decisions of choosing the linkage criterium (*e.g. single linkage, complete linkage, etc.*) as well as the distance function (*Euclidean, Manhattan, etc.*).

We experimented with many different combinations, even with a custom distance function that applied a heavier weight to selected indicators. Ultimately, the best performing one was a *Euclidean* distance metric and the *Ward method* as the linkage criterium. To determine the optimal number of clusters, we performed the gap test just like in k-means, and corroborated our results by graphing and analyzing a dendrogram, which is extremely useful for gaining insight into the "hierarchies" of the clusters. The dendrogram we produced in our run is shown in Figure 5-1. Because hierarchical

agglomerative clustering utilizes a bottom-up approach, meaning we start out with each city as a singleton cluster and iteratively merge clusters, each horizontal line represents a merge. Additionally, we performed the gap test, which can be seen in Figure 5-2. From this graph, the best two possible candidates for optimal number of clusters is five or seven. Ultimately, we decided to go with seven for a higher degree of resolution within the clusters.

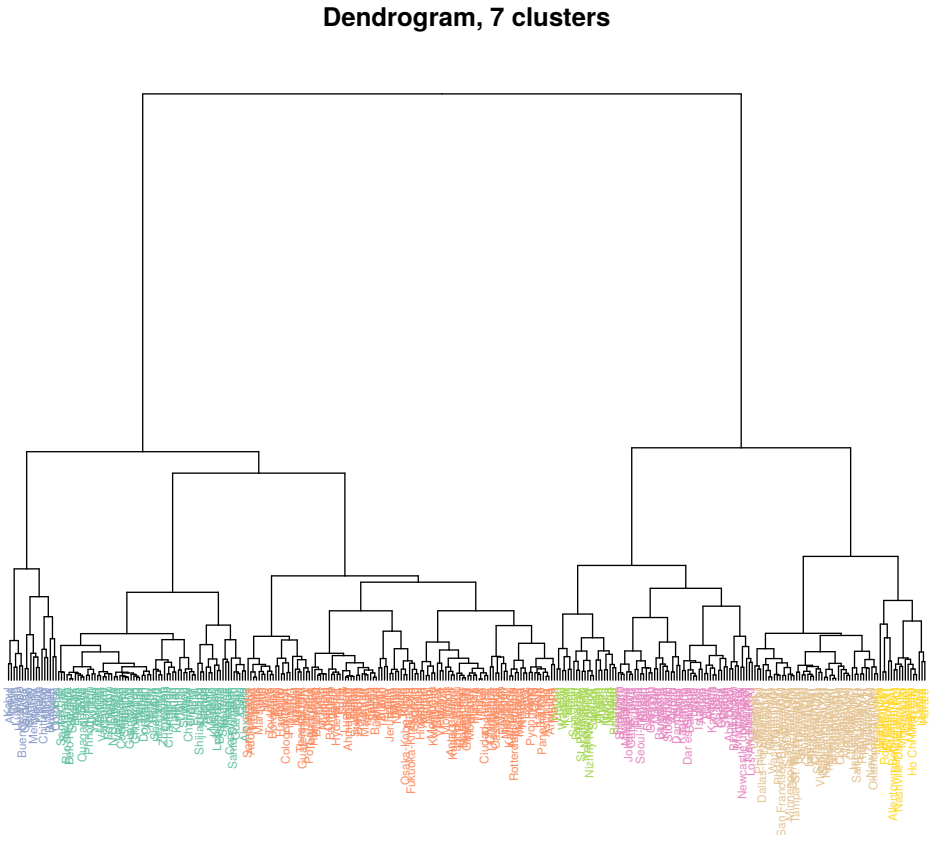


Figure 5-1: Dendrogram from hierarchical agglomerative clustering

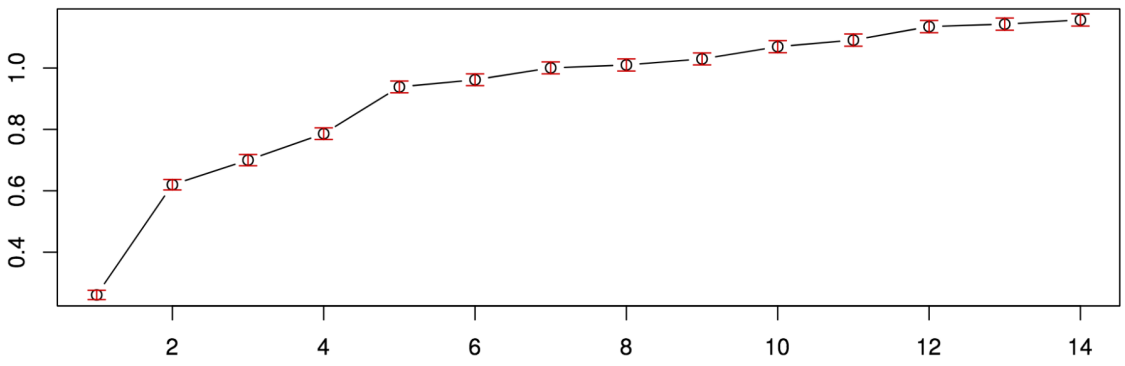


Figure 5-2: Gap test

Chapter 6

Results and Analysis

6.1 Cluster Results

The cluster memberships are shown in listed in Figure 6-1. For a list of the cities in each cluster typology, refer to Table 6.1. Looking at these results, most of the cities that are grouped together are intuitively similar, which is a good sign. For example, it makes sense that Beijing, Bangkok, Mexico City, and Rio are in the same cluster by virtue of all being large, congested, and emerging cities. On the other hand, Boston, San Francisco, and New York, all innovative and car-dominated cities, are grouped together.

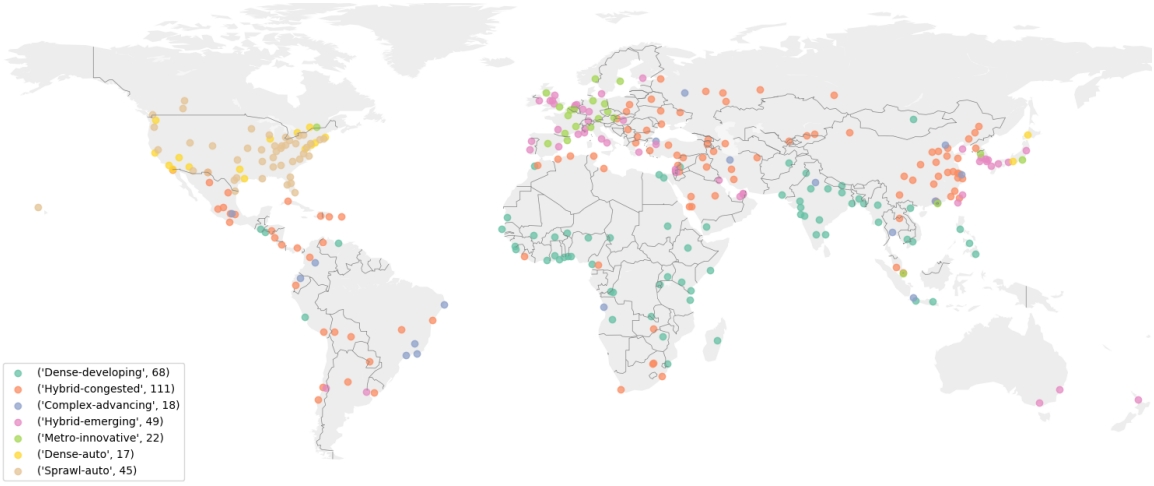


Figure 6-1: Cluster Map

	Typology	#	City Names
1	Dense-developing	69	Abidjan, Accra, Addis Ababa, Ahmedabad, Alexandria, Antananarivo, Bamako, Bandung, Bangalore, Belgrade, Brazzaville, Cairo, Caracas, Casablanca, Cebu, Chennai, Chittagong, Conakry, Cotonou, Dakar, Damascus, Dar es Salaam, Davao, Dhaka, Douala, Freetown, Hanoi, Harare, Ho Chi Minh City, Huambo, Hyderabad, Istanbul, Jaipur, Jakarta, Kabul, Kampala, Kano, Karachi, Kathmandu, Khartoum, Kigali, Kinshasa, Kolkata, Kumasi, Lagos, Lahore, Lome, Lubumbashi, Lucknow, Mandalay, Manila, Maputo, Mogadishu, Mombasa, Mumbai, N'Djamena, Nairobi, Niamey, Nouakchott, Ouagadougou, Patna, Phnom Penh, Pune, Rangoon, Recife, Sanaa, Surabaya, Surat, Ulaanbaatar
2	Hybrid-advancing	87	Aleppo, Algiers, Almaty, Amman, Ankara, Arequipa, Asuncion, Athens, Baku, Beirut, Bishkek, Bratislava, Bucharest, Changchun, Changsha, Chengdu, Chongqing, Dalian, Dongguan, Durban, Fuzhou, Guadalajara, Guatemala City, Hangzhou, Harbin, Havana, Hefei, Izmir, Jeddah, Johor Bahru, Kaohsiung, Kazan, Kharkiv, Kiev, Krakow, Kuala Lumpur, La Paz, Lima, Lusaka, Managua, Maracaibo, Mashhad, Mecca, Medina, Minsk, Monrovia, Mosul, Nanjing, Ningbo, Nizhny Novgorod, Novosibirsk, Odessa, Oran, Panama City, Port-au-Prince, Pyongyang, Qingdao, Rabat, Riyadh, Samara, San Jose, San Salvador, Santa Cruz, Shenyang, Shijiazhuang, Shiraz, Sofia, St. Petersburg, Suzhou, Taiyuan, Tashkent, Tbilisi, Tegucigalpa, Thessaloniki, Tianjin, Tijuana, Toluca, Tripoli, Tunis, Vientiane, Warsaw, Wuhan, Wuxi, Xi'an, Yaounde, Yekaterinburg, Yerevan
3	Congested-emerging	11	Bangkok, Beijing, Delhi, Guangzhou, Luanda, Mexico City, Moscow, Rio de Janeiro, Shanghai, Shenzhen, Tehran
4	Hybrid-innovative	70	Abu Dhabi, Amsterdam, Antwerp, Auckland, Barcelona, Berlin, Birmingham, Bordeaux, Brussels, Budapest, Buenos Aires, Bursa, Busan, Cologne-Bonn, Copenhagen, Daegu, Daejeon, Dubai, Dublin, Frankfurt, Fukuoka-Kitakyushu, Glasgow, Gwangju, Haifa, Hamburg, Helsinki, Hiroshima, Hong Kong, Jerusalem, Kuwait City, Lille, Lisbon, Liverpool, London, Lyon, Madrid, Manchester, Marseille, Melbourne, Milan, Montreal, Munich, Naples, Newcastle upon Tyne, Nice, Osaka-Kobe-Kyoto, Oslo, Paris, Porto, Prague, Rome, Rotterdam-Hague, San Juan, Santiago, Sendai, Seoul-Incheon, Sharjah, Singapore, Stockholm, Sydney, Taipei, Tel Aviv, Tokyo, Toulouse, Turin, Ulsan, Valencia, Valparaiso, Vienna, Zurich
5	BRT-dense	31	Acapulco, Adana, Baghdad, Belo Horizonte, Bogota, Brasilia, Cape Town, Chihuahua, Ciudad Juarez, Concepcion, Cordoba, Guayaquil, Isfahan, Jinan, Johannesburg, Kunming, Leon, Medellin, Monterrey, Montevideo, Pretoria, Puebla, Quito, Salvador, Santo Domingo, Sao Paulo, Tabriz, Taichung, Urumqi, Xiamen, Zhengzhou
6	Auto-innovative	14	Boston(MA), Chicago(IL), Dallas-Fort Worth(TX), Houston(TX), Los Angeles(CA), Nagoya, New York(NY), Philadelphia(PA), San Diego(CA), San Francisco Bay Area(CA), Sapporo, Seattle(WA), Toronto, Washington(DC)
7	Auto-sprawl	48	Allentown-Bethlehem(PA), Atlanta(GA), Austin(TX), Baltimore(MD), Birmingham(AL), Buffalo(NY), Calgary, Charlotte(NC), Cincinnati(OH), Cleveland(OH), Columbus(OH), Dayton(OH), Denver-Aurora(CO), Detroit(MI), Edmonton, El Paso(TX), Hartford(CT), Honolulu(HI), Indianapolis(IN), Jacksonville(FL), Kansas City(MO), Las Vegas(NV), Louisville(KY), McAllen(TX), Memphis(TN), Miami(FL), Milwaukee(WI), Minneapolis-St. Paul(MN), Nashville-Davidson(TN), New Orleans(LA), Oklahoma City(OK), Orlando(FL), Ottawa, Phoenix-Mesa(AZ), Pittsburgh(PA), Portland(OR), Providence(RI), Raleigh(NC), Richmond(VA), Rochester(NY), Sacramento(CA), Salt Lake City(UT), San Antonio(TX), St. Louis(MO), Tampa-St. Petersburg(FL), Tucson(AZ), Vancouver, Virginia Beach(VA)

Table 6.1: Cluster Memberships

6.2 Cluster Typologies

Figure 6-2 shows the average factor values for each cluster. This was very useful in helping coming up with the individual typologies for each cluster, which will be described below:

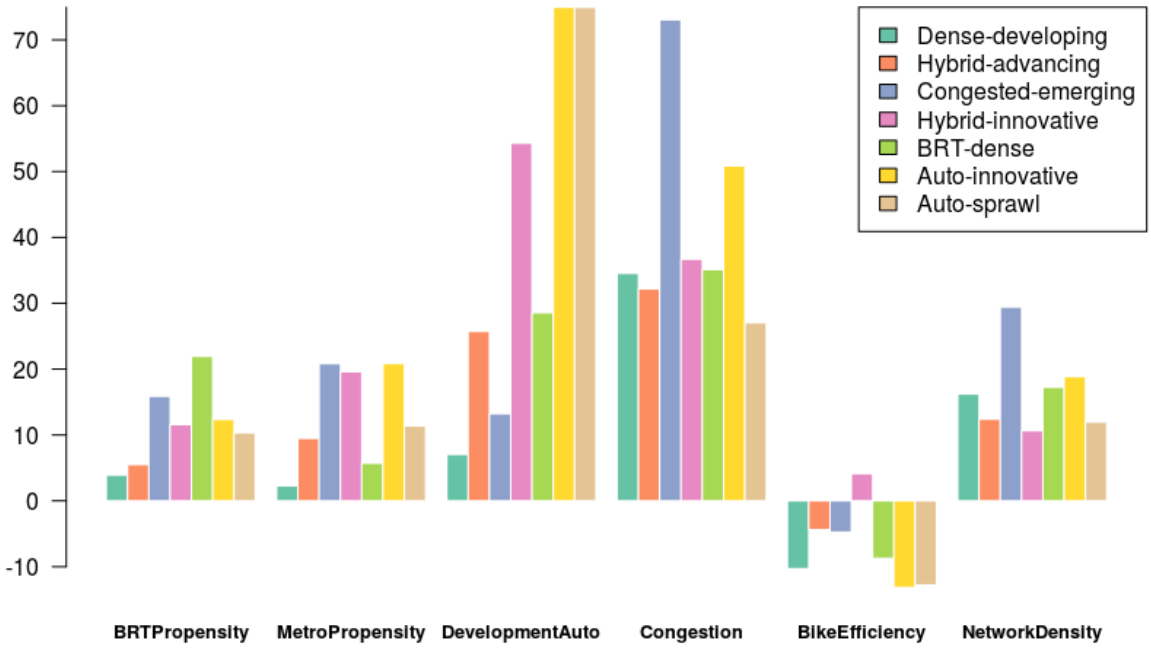


Figure 6-2: Average factor values for each cluster

- **Dense-developing:** The dense-developing type is largely located in Africa and the Indian subcontinent. These are large and growing urban centers, with relatively low development indices. They also have among the highest public transit modeshares, but little availability of metro rail. Of the 7, this typology has the second highest congestion levels and number of intersections.

- **Hybrid-advancing:** This typology is marked by cities with relatively low car use and high pedestrian activity similar to those in Dense-developing) but with greater industrialization and development (in constrast with Dense-developing, which has the lowest aggregate on that factor). They are also moderately sized, and have a metro propensity similar to the Sprawl-auto group. Dominant in this group are cities in North Africa, Eastern Europe, China and Central America.

- **Congested-emerging:** These cities are the highest ranking in congestion compared to those in other groups. However, they rank among the lowest in development and industrialization, hence the term "congested-emerging". Notable in this typology are Beijing, Delhi and Rio de Janeiro. They are also marked by a high BRT propensity, and moderate metro propensity. They have the densest networks as well.

- **Hybrid-innovative:** These cities have a good mix of BRT and metro propensities and rank higher in development than emerging cities. Further, they rank highest on the BikeEfficiency factor, and have the second highest innovation score on average. This typoloy is dominated by European cities as well as leading centers in East Asia (e.g./, Singapore, Sydney, Tokyo) and the Middle East (e.g./, Tel Aviv, Dubai), where urban rail and bikeshare are in abundant supply, where the potential for future mobility solutions is high.

- **BRT-dense:** This group consists of cities with the highest BRTPropensity aggregate. Their development indices are moderate, and they have the lowest MetroPropensity ranking after the Dense-developing typology. The cities of this type are largely found in Latin America (e.g. Concepcion, Sao Paulo, Monterrey, Montevideo) and South Africa, with a few in China and the Middle East.

- **Auto-innovative** These cities are modern and highly industrialized, but marked by a history of automobile-driven development. However, they have the highest average innovation score and digital access index compared to the other typologies. Thus, this typology comprises the subset of relatively dense North American agglomerations with extensive metro transit systems (e.g. Boston,Toronto, New York) along with similarly influenced cities in Japan (Nagoya, Sapporo).

- **Auto-innovative** These cities are modern and highly industrialized, but marked by a history of automobile-driven development. However, they have the highest average innovation score and digital access index compared to the other typologies. Thus, this typology comprises the subset of relatively dense North American

agglomerations with extensive metro transit systems (e.g. Boston,Toronto, New York) along with similarly influenced cities in Japan (Nagoya, Sapporo).

- **Auto-sprawl** This group is made up of the counterpart to the innovative North American city. It ranks lowest on the Congestion factor, and joint-lowest on the BikeEfficiency factor. While the cities of this type share similar car modeshares with those in the Auto-innovative group (86-87%), they have the highest car ownership. Further, they are the least dense in terms of both population and network (intersections). A few examples are Baltimore, Indianapolis and St. Louis.

6.3 Confirmatory Analysis

In addition to the empirical evidence supporting the quality of the clusters, another student in our team has also worked on confirming the results of the clustering, but through more rigorous methods. She created a *Latent Class Choice Model*, which is simply a finite mixture modelat its core. In this context, after creating a set of latent classes, each city is assigned a certain probability to belong to a class. This can be useful in corroborating that members of a certain cluster indeed do have a decent probability of belonging in that class according to the model [26].

To achieve these latent classes, the student not only used our dataset, but also incorporated a behavioral dataset that was helped obtained by a service called *Dalia Research* which surveyed thousands of people in cities around the world. The questions on the survey were binary, and generally asked transportation-related questions like "Do you own a car?" or "Is owning a car a symbol of status in your city?". Final results partially confirmed the clusters that were produced from my work, but the Latent Class Choice Model seemed better at grouping together similar cities from different countries, which perhaps shows that the behavioral data has a significant influence on defining a city.

6.4 Transportation Research Board Conference

The results and analysis shown here have also been included in a paper to be presented at the 2018 Transportation Research Board Meeting [26], which is an exciting platform to show the world the innovative work we have done on urban city typologies. The conclusions that we have reached are very important in the context of future mobility and have the potential to greatly influence how organizations around the world react to inevitable changes in transportation patterns and policies.

Chapter 7

Dashboard

The clustering results are a prerequisite for determining and describing the prototype cities necessary for fitting our models of future mobility simulation, but they are not just a means to an end. In fact, there are many interesting conclusions one can draw by analyzing the cluster memberships and the defining characteristics of each group. Thus, we thought it would be extremely useful to create a dashboard so that others can not only view the clusters in an easily understandable format but also customize and visualize graphs of the city data.

7.1 Platform

The unequivocally best platform to host the dashboard would be a web app, and that is the direction I went with from the get-go. The advantages are very clear: anyone can access it on the Internet simply through a URL, and patches can be applied at any time to improve or update the interface and contents. Subsequently, the application was built mostly with HTML/CSS and Javascript, all languages that I had prior experience in.

7.1.1 D3.js

D3.js [1] is one of the most popular web frameworks for "manipulating documents based on data" and is an extremely powerful tool for visualizing data. Although it has a notorious reputation for having a steep learning curve, I had previously worked with the library and was confident in my ability to utilize it for the dashboard. There are also many open-source tools built upon D3 that specialize in a certain area and add more functionality there. For the dashboard, I personally took advantage of *NVD3* [5], which is an excellent library for designing and creating dynamic charts and graphs, and *DataMaps* [2], a great resource for making dynamic maps.

7.2 Web App Interface

My vision for the interface was to have multiple pages that are easily navigable from a menu, each offering something different about the project. The contents could contain any number of interesting items, but I decided that the *home page* should be a visually stunning and easy-to-understand interface that would present a holistic perspective of the project, and what better than to have a dynamic world map. The other pages should allow the viewer to learn more about the cities and resources/indicators involved in the dataset, as well as be able to visualize the raw data that we collected. I will attempt to give a brief tour of the dashboard with supporting images, but for a genuine experience, feel free to explore at `http://web.mit.edu/its-lab/www/dashboard/map.html`.

7.2.1 Home Page: Cluster Map

The world map is meant to be the crown jewel of the dashboard, and it should immediately grab the viewer's attention. As one can see in Figure 7-1, the map not only contains many differently sized and colored bubbles, but also an explanatory legend and circular menu selectors, as well as a *Reset* button. The legend matches each color to its corresponding cluster typology, and one can easily consult it to determine which cluster any city belongs to. Additionally, interacting with the legend

allows filtering based on the active bubbles, so users are able to view only cities of selected clusters. The map fully supports panning and zooming to get a closer look at the cities, and hitting the *Reset* button transforms the map back to its original state.

The "menu" on the bottom left corner currently allows users to select among four different indicators — population, CO_2 emissions per capita, motor vehicles per capita, and GDP per capita. We plan to extend the number of options in the future, but this is what we have so far. The size of the bubble over a city is proportional to the value of the indicator for that particular city. Clicking on any of the buttons on the will switch the active selected indicator, and that will be reflected in the changing of the bubble sizes. Hovering over these bubbles will trigger a description box to appear, listing the name of the city as well as the value of the selected indicator. All of these functions are intuitive and easy to use, providing a very powerful yet simple interface.

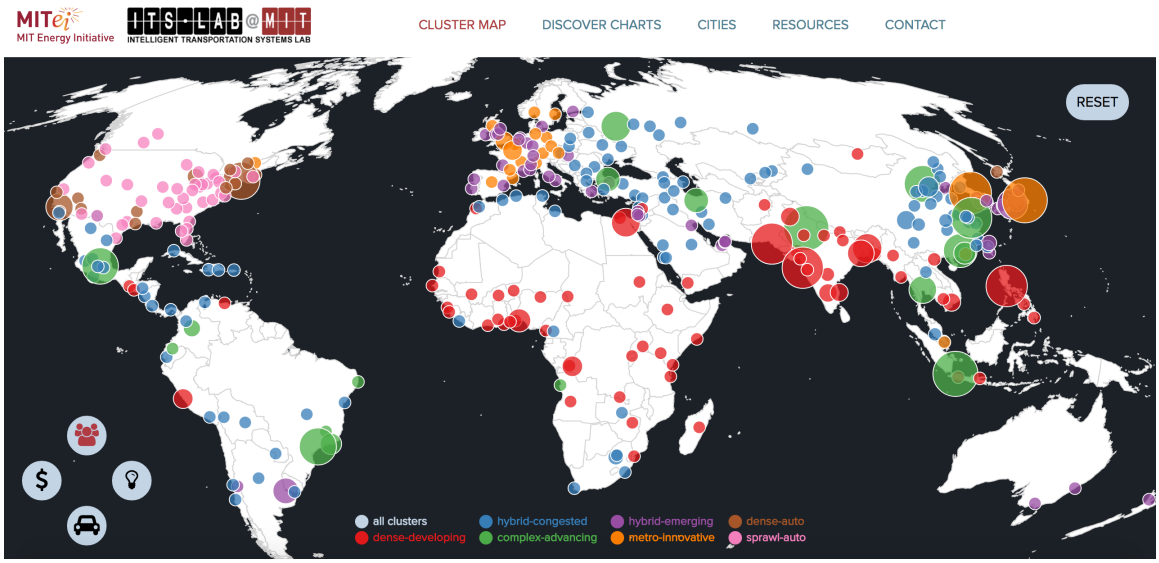


Figure 7-1: Dynamic world map with clusters

7.2.2 Cities

This page in the dashboard is meant to highlight one city at a time. The client can use the search bar (with auto-complete support) to easily look for a particular city of interest. Once that is done, a list of all the data that was collected for the city will be

shown, with the corresponding indicator values (Figure 7-2). At the bottom of any city page is a complete list of all the cities in our dataset, sorted by cluster typology (Figure 7-3). This is especially useful for people who simply want to explore and learn more about the cities in a specific cluster.

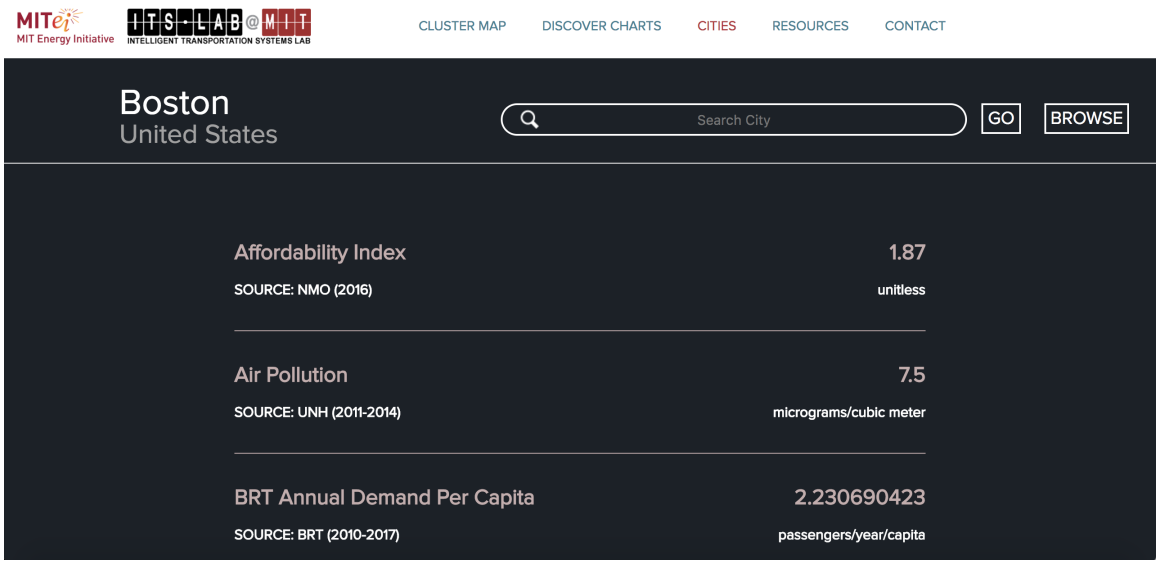


Figure 7-2: Search for city to view

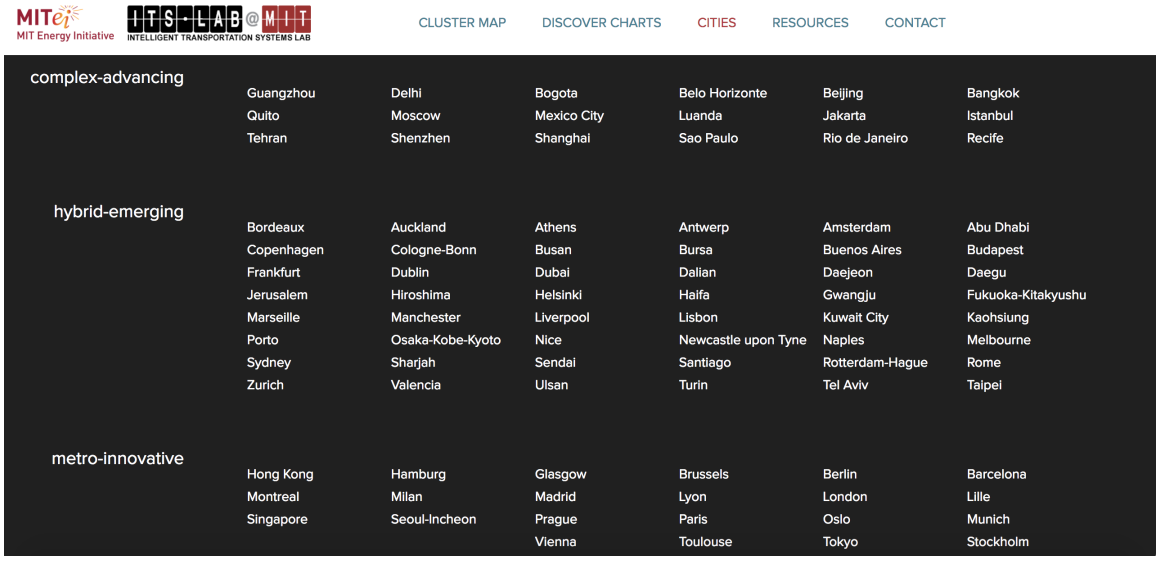


Figure 7-3: Selection of cities by cluster

7.2.3 Discover Charts

Data is best understood when it is visualized, motivating the creation of this page. Ideally there would be many different kinds of graphs that clients can view and play around with, but there are currently two types being offered — average factor values for each cluster (Figure 7-4), and customizable scatter plots for indicators (Figure 7-5).

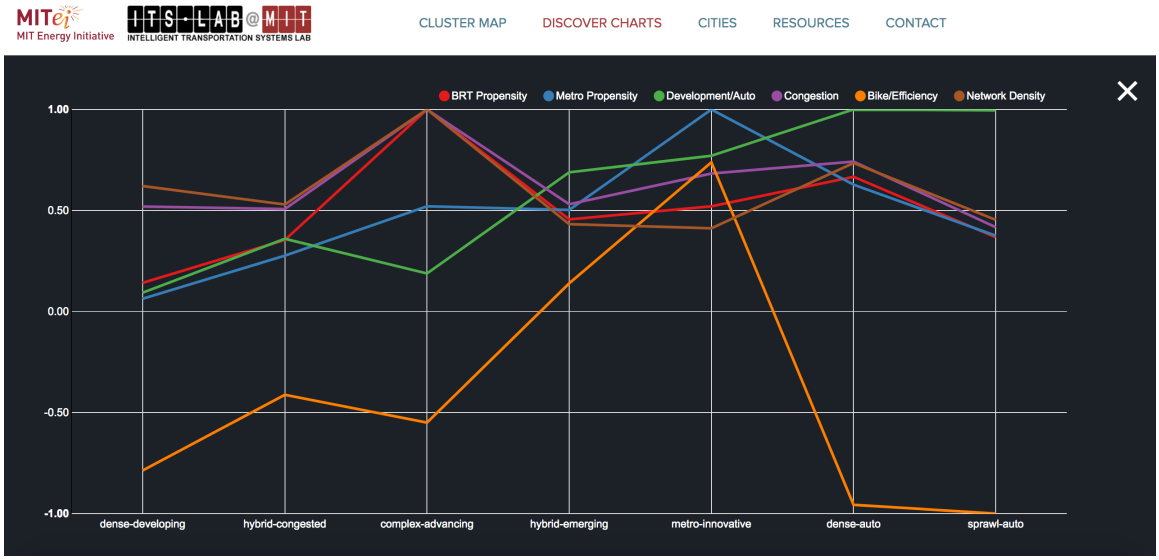


Figure 7-4: Plot of average factor scores for clusters

Figure 7-5: Customizable scatter plot

The cluster factor average graph shows how the factor values vary for each cluster typology, allowing users to easily compare and contrast among the clusters, which are colored appropriately. The interface has support for toggling factors on/off, so one can pick and choose any combination of factors to be displayed. Take note that the factor scores have all been linearly transformed to fit within a range of [-1,1] in order to avoid being potentially visually misleading.

The customizable scatter plots offer a direct comparison between two indicators for every city, each of which is colored according to its cluster typology. Once again, the interface has support for picking and choosing which clusters to be displayed. This graph is especially useful for not only observing whether or not there is correlation between two indicators that are seemingly related, but also corroborating the cluster memberships by checking if cities of the same color are grouped together on the chart. There are two search bars (one for each axis), and auto-complete functionality is provided for ease of use.

7.2.4 Miscellaneous

For logistical purposes, there is a *Contact* page, so clients can reach our team and provide potential feedback. There is also a list of sponsors to show which organizations are interested and invested in our work. Also under construction is a *Resources* page, which is intended to elucidate the indicators we collected and which sources they came from. Once again, a huge benefit of hosting this dashboard online is the privilege of patching it anytime and having the update propogate through the web immediately. We plan on putting the finishing touches soon and hope to add many more available graphs in the future!

Chapter 8

Future Work

Recall from the very beginning that the objective of the *Mobility of the Future* project is to create models simulating future urban mobility patterns and different transportation energy scenarios. While what has been done so far (up to achieving the city clusters) is imperative, there is still considerable work to be done in order to fulfill the goals. Although I unfortunately will not be present to see it through, I am excited for those who will be involved till the end, as the subsequent steps of prototype city generation and incorporation of the SimMobility [6] platform are technically challenging but rewarding tasks.

8.1 Prototype City Generation

With our final clusters determined and analyzed, there is enough information to produce a prototype city for each group. The proposed plan is to generate virtual cities using the defining characteristics of each cluster. The city model that SimMobility takes as its input has two components — *demand* and *road network*. Thus, the main focus while generating the virtual cities will be to obtain accurate representations of those for each cluster.

To appropriately model the demand, we must first get a sense of the demographics of the city. This means understanding how the population is allocated, what an average

household comprises, and how different attributes related to age, sex, education, etc. can affect people's everyday travel patterns. There has been substantial prior work studying how to properly model travel demand, and potential methodologies and algorithms that we will consider include *Iterative Proportional Fitting* [20], a *Bayesian network* approach [38], and *Markov Chain Monte Carlo - based sampling* [25].

To generate an accurate road network model, we are fortunate to have an expansive set of network statistics from OpenStreetMap to consult. By basing our model on the network structures of each cluster, we can ensure the creation of an appropriate model. One person on our team has already made an attempt at this. Taking into account desired network statistics, he seeded initial node locations on a blank city canvas, then algorithmically connected them to create streets, ultimately producing an entire road network.

In addition, there does exist some literature that could prove helpful, including work done by Dai et. al [24], but surprisingly, what could be truly edifying is studying how the engine of city simulation games work. The classic example is obviously *SimCity*, but a better template is the more recent *Cities: Skyline*, which runs on quite an advanced engine and has an impressively accurate model for traffic and road networks. The generation of these virtual cities is going to be the most challenging task that remains, but should definitely be a fun and exciting journey.

8.2 SimMobility

SimMobility is the "simulation platform of the Future Urban Mobility Research Group at the Singapore-MIT Alliance for Research and Technology (SMART) that aims to serve as the nexus of Future Mobility research evaluations". From what I understand, it is the default tool that will be used to simulate future mobility scenarios on the generated prototype cities. Given that many members of our team are actually actively involved in SimMobility, setup, execution, and troubleshooting of the tool will be

relatively easy, thus providing a huge in-house advantage. Given the complexity of the algorithms and the size of the virtual cities, a successful run will take a significant amount of time, but once all the results are produced, the only remaining step is to perform analysis and assess the future landscape of mobility.

Chapter 9

Conclusion

The *Mobility of the Future* project hosted by the MIT Intelligent Transportation Systems Lab and sponsored by the MIT Energy Initiative is a mission motivated by a need to properly understand how the future of global urban transportation will be affected by inevitable changes in demographics, economy, and energy regulations. To capture the global scope while maintaining a manageable scale is tricky, but can be done if we accurately cluster the cities around the world into groups, each with its own unique defining characteristics. My year-long role in this task was to carry out this vision.

Many design decisions were made prior to making any definitive decisions, including determining which cities to include and what kind of data to collect. Over the course of several months, the grunt work of populating a dataset with relevant information was done, ultimately culminating in a CSV file with nearly 100 indicators for over 300 cities. Before diving headfirst into the clustering process, we still needed to do some prep work, including filling the necessary holes in data and performing factor analysis to reduce dimensions down to seven defining factors.

With a much more practical transformed dataset in place, we then proceeded to test a few different clustering methods, including K-means and several variants of hierarchical clustering. Ultimately, we found hierarchical clustering using the Ward

method to be most effective, resulting in a final tally of seven clusters. After performing the appropriate analysis on the characteristic values for each cluster as well as the composition of the defining factors, we were able to not only provide descriptions for the factors, but also introduce typological nomenclature for the clusters.

To present the results in a more understandable manner, we designed and created a web dashboard to visualize the data using a world map marked by bubbles located at cities in our dataset, colored appropriately to denote what cluster they are in. There is also a legend matching color to cluster typology, allowing users to easily cross-reference between the two. The dashboard has more useful functionalities, including viewing the raw data we collected for each city as well as customizing and visualizing custom graphs based on the data.

While my contributions to the project end here, *Mobility of the Future* still requires more work to be done, and I am excited to the see the conclusion and view the results when others complete the task. Overall, I feel blessed to be given this opportunity to apply my area of expertise in computer science to a completely different realm in civil engineering. It has truly been a humbling and edifying experience, as I have learned so much about not only transportation and cities in general, but also about machine learning, data analysis, and even web programming and design!

Appendix A

Cities in the Dataset

Table A.1: List of every city in our dataset, sorted by country

Country	Cities	Count
Afghanistan	Kabul	1
Algeria	Algiers, Oran	2
Angola	Luanda, Huambo	2
Argentina	Buenos Aires, Cordoba	2
Armenia	Yerevan	1
Australia	Sydney, Melbourne	2
Austria	Vienna	1
Azerbaijan	Baku	1
Bangladesh	Dhaka, Chittagong	2
Belarus	Minsk	1
Belgium	Brussels, Antwerp	2
Benin	Cotonou	1
Bolivia	La Paz, Santa Cruz	2
Brazil	Sao Paulo, Rio de Janeiro, Belo Horizonte, Salvador, Brasilia, Recife	6
Bulgaria	Sofia	1
Burkina Faso	Ouagadougou	1
Cambodia	Phnom Penh	1
Cameroon	Douala, Yaounde	2
Canada	Toronto, Montreal, Vancouver, Calgary, Edmonton, Ottawa	6
Chad	N'Djamena	1
Chile	Santiago, Valparaiso, Concepcion	3
China	Shanghai, Beijing, Guangzhou, Shenzhen, Chongqing, Wuhan, Tianjin, Dongguan, Chengdu, Nanjing, Harbin, Shenyang, Hangzhou, Xi'an, Zhengzhou, Qingdao, Changchun, Jinan, Taiyuan, Kunming, Dalian, Suzhou, Wuxi, Changsha, Urumqi, Hefei, Fuzhou, Shijiazhuang, Xiamen, Ningbo	30
Colombia	Bogota, Medellin	2
Congo	Brazzaville	1
Costa Rica	San Jose	1
Cuba	Havana	1
Czech Republic	Prague	1
Democratic Republic of the Congo	Kinshasa, Lubumbashi	2
Denmark	Copenhagen	1
Dominican Republic	Santo Domingo	1

Country	Cities	Count
Ecuador	Guayaquil, Quito	2
Egypt	Cairo, Alexandria	2
El Salvador	San Salvador	1
Ethiopia	Addis Ababa	1
Finland	Helsinki	1
France	Paris, Marseille, Lyon, Lille, Nice, Toulouse, Bordeaux	7
Georgia	Tbilisi	1
Germany	Berlin, Hamburg, Munich, Cologne-Bonn, Frankfurt	5
Ghana	Accra, Kumasi	2
Greece	Athens, Thessaloniki	2
Guatemala	Guatemala City	1
Guinea	Conakry	1
Haiti	Port-au-Prince	1
Honduras	Tegucigalpa	1
Hong Kong	Hong Kong	1
Hungary	Budapest	1
India	Delhi, Mumbai, Kolkata, Chennai, Bangalore, Jaipur, Hyderabad, Pune, Surat, Lucknow, Ahmedabad, Patna	12
Indonesia	Jakarta, Surabaya, Bandung	3
Iran	Tehran, Mashhad, Isfahan, Tabriz, Shiraz	5
Iraq	Baghdad, Mosul	2
Ireland	Dublin	1
Israel	Tel Aviv, Haifa, Jerusalem	3
Italy	Rome, Milan, Naples, Turin	4
Ivory Coast	Abidjan	1
Japan	Tokyo, Osaka-Kobe-Kyoto, Nagoya, Fukuoka-Kitakyushu, Sapporo, Sendai, Hiroshima	7
Jordan	Amman	1
Kazakhstan	Almaty	1
Kenya	Nairobi, Mombasa	2
Kuwait	Kuwait City	1
Kyrgyzstan	Bishkek	1
Laos	Vientiane	1
Lebanon	Beirut	1
Liberia	Monrovia	1
Libya	Tripoli	1
Madagascar	Antananarivo	1
Malaysia	Kuala Lumpur, Johor Bahru	2
Mali	Bamako	1
Mauritania	Nouakchott	1
Mexico	Mexico City, Guadalajara, Monterrey, Puebla, Tijuana, Toluca, Leon, Ciudad Juarez, Acapulco, Chihuahua	10
Mongolia	Ulaanbaatar	1
Morocco	Casablanca, Rabat	2
Mozambique	Maputo	1
Myanmar	Rangoon, Mandalay	2
Nepal	Kathmandu	1
Netherlands	Amsterdam, Rotterdam-Hague	2
New Zealand	Auckland	1
Nicaragua	Managua	1
Niger	Niamey	1
Nigeria	Lagos, Kano	2
North Korea	Pyongyang	1
Norway	Oslo	1
Pakistan	Karachi, Lahore	2
Panama	Panama City	1
Paraguay	Asuncion	1
Peru	Lima, Arequipa	2
Philippines	Manila, Davao, Cebu	3

Country	Cities	Count
Poland	Warsaw, Krakow	2
Portugal	Lisbon, Porto	2
Puerto Rico	San Juan	1
Romania	Bucharest	1
Russia	Moscow, St. Petersburg, Novosibirsk, Yekaterinburg, Nizhny Novgorod, Samara, Kazan	7
Rwanda	Kigali	1
Saudi Arabia	Riyadh, Mecca, Medina, Jeddah	4
Senegal	Dakar	1
Serbia	Belgrade	1
Sierra Leone	Freetown	1
Singapore	Singapore	1
Slovakia	Bratislava	1
Somalia	Mogadishu	1
South Africa	Johannesburg, Cape Town, Durban, Pretoria	4
South Korea	Seoul-Incheon, Busan, Daegu, Daejeon, Gwangju, Ulsan	6
Spain	Madrid, Barcelona, Valencia	3
Spain	Madrid, Barcelona, Valencia	3
Sudan	Khartoum	1
Sweden	Stockholm	1
Switzerland	Zurich	1
Syria	Aleppo, Damascus	2
Taiwan	Taipei, Kaohsiung, Taichung	3
Tanzania	Dar es Salaam	1
Thailand	Bangkok	1
Togo	Lome	1
Tunisia	Tunis	1
Turkey	Istanbul, Ankara, Izmir, Bursa, Adana	5
Uganda	Kampala	1
Ukraine	Kiev, Kharkiv, Odessa	3
United Arab Emirates	Dubai, Sharjah, Abu Dhabi	3
United Kingdom	London, Birmingham, Manchester, Glasgow, Newcastle upon Tyne, Liverpool	6
United States	New York(NY), Los Angeles(CA), Chicago(IL), Miami(FL), Philadelphia(PA), Dallas-Fort Worth(TX), Atlanta(GA), Houston(TX), Boston(MA), Washington(DC), Detroit(MI), Phoenix-Mesa(AZ), San Francisco Bay Area(CA), Seattle(WA), San Diego(CA), Minneapolis-St. Paul(MN), Denver-Aurora(CO), Tampa-St. Petersburg(FL), Baltimore(MD), St. Louis(MO), Portland(OR), Cleveland(OH), Las Vegas(NV), Pittsburgh(PA), Cincinnati(OH), Sacramento(CA), Virginia Beach(VA), San Antonio(TX), Kansas City(MO), Indianapolis(IN), Milwaukee(WI), Orlando(FL), Providence(RI), Columbus(OH), Austin(TX), Memphis(TN), Buffalo(NY), Charlotte(NC), Jacksonville(FL), Salt Lake City(UT), Louisville(KY), Richmond(VA), Hartford(CT), Tucson(AZ), New Orleans(LA), Oklahoma City(OK), Honolulu(HI), Dayton(OH), McAllen(TX), Rochester(NY), El Paso(TX), Raleigh(NC), Allentown-Bethlehem(PA), Birmingham(AL), Nashville-Davidson(TN)	55
Uruguay	Montevideo	1
Uzbekistan	Tashkent	1
Venezuela	Caracas, Maracaibo	2
Vietnam	Ho Chi Minh City, Hanoi	2
Yemen	Sanaa	1
Zambia	Lusaka	1
Zimbabwe	Harare	1

Appendix B

Sources and Indicators

Table B.1: List of every source and descriptions of its indicators

Source(s)	Years	Indicators	Units	Description
Demographia	2016	Population	people	number of people in metro area
DegreeDays.net (Weather Underground)	2012-16	Heating Degree Days	heating degree days	higher means colder on average
		Cooling Degree Days	cooling degree days	inverse of heating degree days
Global Bus Rapid Transit	2010-17	BRT Fleet Size	buses	total number of buses
		BRT Fare	US dollars	cost of a bus trip
		BRT Stations	stations	total number of bus stations
		BRT System Length	kilometers	length of road coverage by buses
		BRT Annual Ridership	people	number of passengers per year
		BRT Fleet Size Per 100k	buses/100k people	total number of buses per 100k people
		BRT Stations Per 100k	stations/100k people	total number of bus stations per 100k people
		BRT System Length Density	per km	total length divided by metro area
		BRT Annual Ridership Per Capita	passengers/year/capita	total annual ridership divided by metro population
Global City Indicators	2013	GDP Per Capita	US dollars/capita	GDP per capita
		Poverty Rate	percentage	percentage of population in poverty
		Infant Mortality	percentage	percentage of children who die before age 1
		Life Expectancy	years	average time expected to live
Global Petrol Prices	2017	Gasoline Price	US dollars	price of gas
Innovation Cities Index	2015	Innovation Score	-	estimation of innovation: [0-100] scale (higher = more innovative)
Internet World Stats	2017	Internet Penetration	percentage	percentage of population with Internet access
		Digital Access Index	-	0-1 scale (higher is better)
Numbeo	2016	Cost of Living Index	percentage	percentage w.r.t New York City
		Rent Index	percentage	percentage w.r.t New York City
		Groceries Index	percentage	percentage w.r.t New York City
		Restaurant Price Index	percentage	percentage w.r.t New York City
		Local Purchasing Power Index	percentage	percentage w.r.t New York City
		Price To Income Ratio	ratio	ratio of median apartment prices to median annual family income

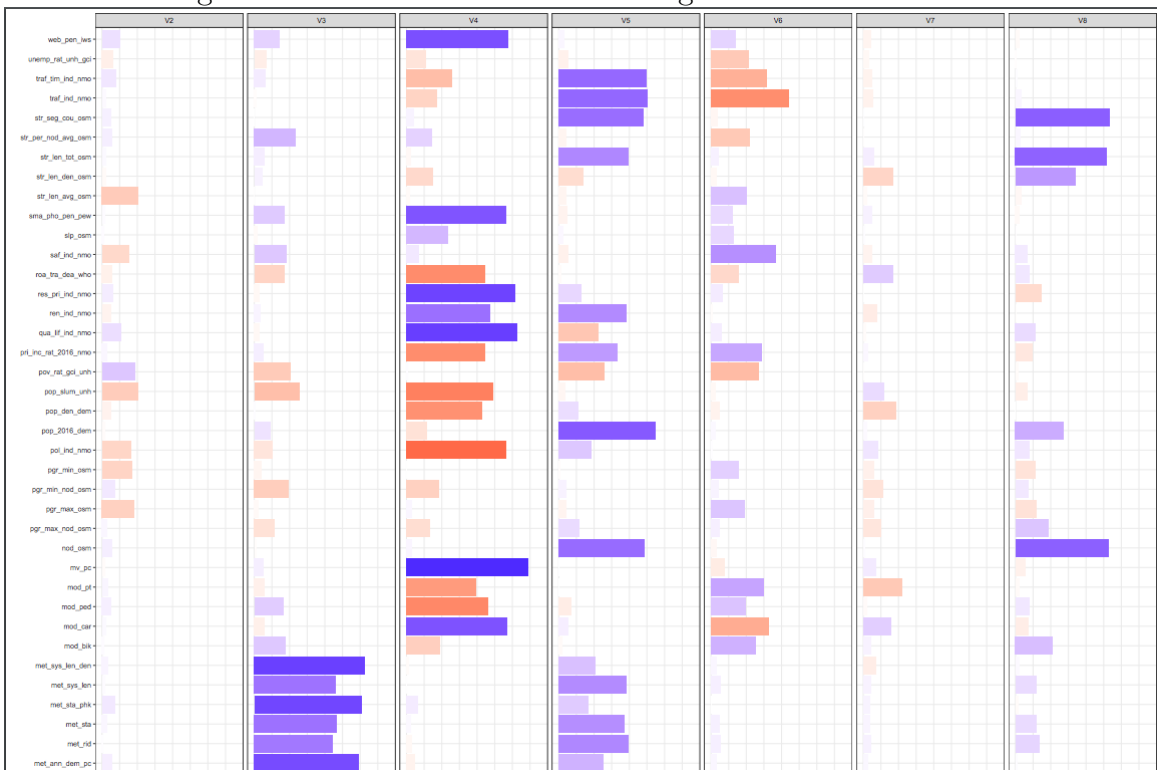
Source(s)	Years	Indicators	Units	Description
Numbeo	2016	Affordability Index	-	inverse of mortgage as percentage of income
		Crime Index	-	0-100 scale
		Safety Index	-	0-100 scale
		Health Care Index	-	0-100 scale
		Pollution Index	-	0-100 scale
		Traffic Index	-	composite index of time consumed in traffic due to job commute, estimation of time consumption dissatisfaction, CO_2 consumption estimation in traffic and overall inefficiencies in the traffic system
		Traffic Time Index	minutes	average one way time needed to transport
		Inefficiency Index	-	estimation of inefficiencies in the traffic
		CO_2 Emission Index	grams	estimation of CO_2 consumption due to traffic time for commute to and from work
		Quality Of Life Index	-	estimation of overall quality of life
		Climate Index	-	estimation of climate likability: scale of [-100,100]
OpenStreetMap	2017	circuity avg	percentage	edge length total divided by the sum of the great circle distances between the nodes of each edge
		edge length avg	meters	mean edge length in the graph
		street length total	meters	sum of all edge lengths in the undirected representation of the graph
		street length density	per km	total street length divided by metro area
		avg neighbor degree avg	degrees	mean of neighbor degrees for all nodes
		avg weighted neighbor degree avg	degrees	weighted mean of neighbor degrees for all nodes
		degree centrality avg	percentage	fraction of nodes that each node is connected to
		clustering coefficient avg	-	mean of clustering coefficients of all nodes in network
		clustering coefficient weighted avg	-	mean of weighted clustering coefficients of all nodes in network
		pagerank max node	node	node with the maximum pagerank
		pagerank max	pagerank	highest pagerank value of any node in the graph
		pagerank min node	node	node with the minimum pagerank
		pagerank min	pagerank	lowest pagerank value of any node in the graph
		n	nodes	number of nodes in network
		m	nodes	number of edges in network
		k avg	edges	mean number of inbound and outbound edges incident to the nodes
		count intersections	intersections	number of intersections in graph (intersection = node with more than 1 edge emanating from it)
		streets per node avg	streets/node	mean number of physical streets that emanate from each node (intersections and dead-ends)
		edge length total	meters	sum of all edge lengths in the graph
		self loop proportion	percentage	proportion of edges that have a single node as its two endpoints (ie, the edge links nodes u and v, and u=v)
		street length avg	meters	mean edge length in the undirected representation of the graph
		street segments count	streets	number of edges in the undirected representation of the graph
		intersection density	intersections/street	number of intersections divided by total number of streets
Pew Research Center	2017	Smartphone Penetration	percentage	percentage of population with smartphones
Simple Maps	2017	Latitude	degrees	latitude
		Longitude	degrees	longitude

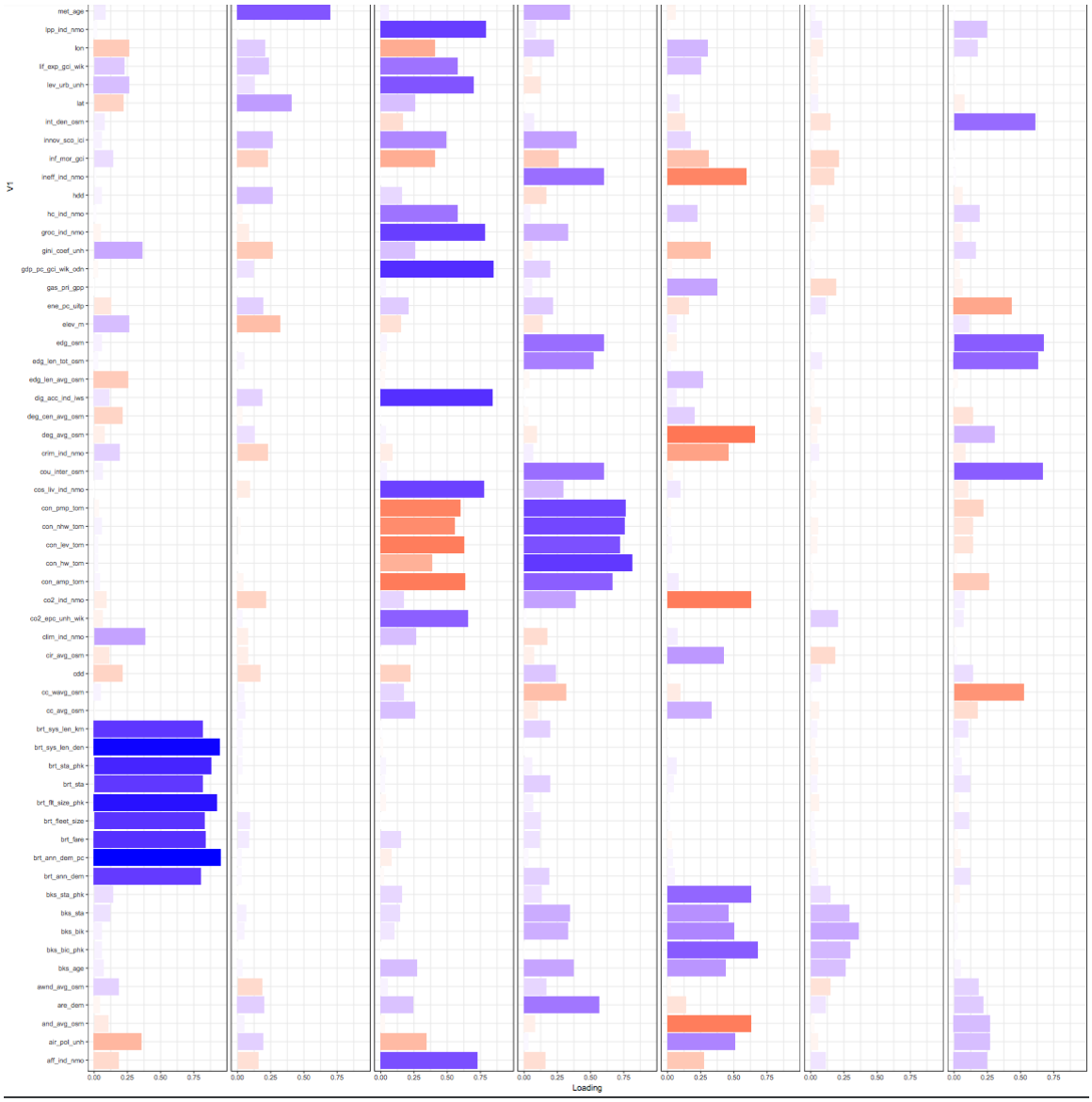
Source(s)	Years	Indicators	Units	Description
TomTom	2016	Congestion Level	percentage	percentage increase in overall travel time compared to free flow situation
		Morning Peak	percentage	percentage increase in morning peak travel time compared to free flow situation
		Evening Peak	percentage	percentage increase in evening peak travel time compared to free flow situation
		Highways	percentage	percentage increase in highway travel time compared to free flow situation
		Non-Highways	percentage	percentage increase in non-highway travel time compared to free flow situation
Union Internationale des Transports Publics	2015	Annual Energy Consumption Per Capita	GJ/year/capita	average amount of annual energy consumed per capita
		Land Area	square kilometers	size of metro area
		Population Density	people/sq. km.	number of people per square kilometer
UN-Habitat	2011-14	Gini Coefficient	-	measure of inequality in wealth distribution: [0-100] scale (lower = more equal)
		CO_2 Emissions Per Capita	tonnes/capita	Amount of CO_2 emissions per capita
		Air Pollution Density	micrograms/cubic meter	Micrograms of air pollution per cubic meter
		Unemployment Rate	percentage	percentage of pollution that is unemployed
		Level of Urbanization	percentage	percentage of metro population in urban areas
		Percentage of Urban Population in Slums	percentage	percentage of urban population in slums
World Health Organization	2013	Annual Road Traffic Deaths Per Capita	deaths/year/100k people	average annual traffic deaths per 100k people
Various Sources (Wikipedia, City Clock, BRT, EPOMM)	2010-17	Car Modeshare	percentage	percentage of population that uses private vehicles for transportation
		Public Transit Modeshare	percentage	percentage of population that uses public transit for transportation
		Bicycle Modeshare	percentage	percentage of population that bikes for transportation
		Walking Modeshare	percentage	percentage of population that walks for transportation
		Years Since Bikeshare Inauguration	years	number of years since bikeshare program started
		Number of Bikeshare Stations	stations	total number of bikshare stations
		Number of Bikeshare Bicycles	bicycles	total number of bikeshare bikes
		Bikeshare Stations Per 100k	stations/100k people	total number of bikeshare stations per 100k people
		Bikeshare Bicycles Per 100k	bicycles/100k people	total number of bikeshare bikes per 100k people
		Elevation	meters	average elevation of metro area
		Years Since Metro Opening	years	number of years since metro opened
		Metro Stations	stations	total number of metro stations
		Metro System Length	kilometers	total length of metro system
		Annual Metro Ridership	people	number of passengers per year
		Metro Stations Per 100k	stations/100k people	number of annual passengers per 100k people
		Metro System Length Density	per km	total system length divided by metro area
		Metro Annual Demand Per Capita	people/year/capita	total annual ridership divided by metro population
		Motor Vehicles Per Capita	vehicles/capita	number of vehicles per capita

Appendix C

Factor Loadings

Figure C-1: Chart of indicator loading scores for each factor




V1
met_age
lpp_ind_nmo
lon
lif_exp_gci_wik
lev_urb_unh
lat
int_den_osm
innov_sco_ici
inf_mor_gci
ineff_ind_nmo
hdd
hc_ind_nmo
groc_ind_nmo
gini_coef_unh
gdp_pc_gci_wik_odn
gas_pri_gpp
ene_pc_ulb
elev_m
edg_osm
edg_len_tot_osm
edg_len_avg_osm
dig_acc_ind_iws
deg_cen_avg_osm
deg_avg_osm
crim_ind_nmo
cou_inter_osm
cos_liv_ind_nmo
con_pmp_tom
con_nhw_tom
con_lev_tom
con_hw_tom
con_amp_tom
co2_ind_nmo
co2_epc_unh_wik
clim_ind_nmo
cir_avg_osm
cdd
cc_waivg_osm
cc_avg_osm
brt_sys_len_km
brt_sys_len_den
brt_sta_phk
brt_sta
brt_flt_size_phk
brt_fleet_size
brt_fare
brt_ann_dem_pc
brt_ann_dem
bks_sta_phk
bks_sta
bks_bik
bks_bic_phk
bks_age
awnd_avg_osm
are_dem
and_avg_osm
air_pol_unh
aff_ind_nmo
0.00
0.25
0.50
0.75
Loading

Bibliography

[1] D3: Data-driven documents. https://d3js.org/.

[2] Datamaps: Customizable svg map visualizations for the web in a single javascript file using d3.js. http://datamaps.github.io.

[3] Degree days (weather underground). http://www.degreedays.net.

[4] Global bus rapid transit data. http://brtdata.org.

[5] Nvd3: re-usable charts for d3.js. http://nvd3.org.

[6] Simmobility - integrated simulation platform. https://its.mit.edu/research/simmobility_v0.

[7] Union internationales des transports publics. http://www.uitp.org.

[8] United nations habitat. http://urbandata.unhabitat.org.

[9] World health organization. http://apps.who.int/gho/data/node.main.A997.

[10] Tomtom. https://www.tomtom.com/en_gb/trafficindex, 2016.

[11] Innovation cities index. http://www.innovation-cities.com/innovation-cities-index-2016-2017-global, 2016-2017.

[12] City proper. https://en.wikipedia.org/wiki/City_proper, Jul 2017.

[13] Demographia. http://www.demographia.com/db-worldua.pdf, 2017.

[14] Global petrol prices. http://www.globalpetrolprices.com/gasoline_prices, 2017.

[15] Internet world stats. http://www.internetworldstats.com, 2017.

[16] New york city global partners. http://www.nyc.gov/html/ia/gprb/html/global/global.shtml, 2017.

[17] Numbeo. https://www.numbeo.com/common, 2017.

[18] Openstreetmap. https://www.openstreetmap.org, 2017.

[19] Simple maps. http://simplemaps.com/data/world-cities, 2017.

[20] Richard J Beckman, Keith A Baggerly, and Michael D McKay. Creating synthetic baseline populations. *ransportation Research Part A: Policy and Practice*, 30(6):415–429, Nov 1996.

[21] Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *SSRN Electronic Journal*.

[22] Pádraig Cunningham. Dimension reduction. *Machine Learning Techniques for Multimedia Cognitive Technologies*, page 91–112.

[23] Ralph B. Dagostino and Heidy K. Russell. Scree test. *Encyclopedia of Biostatistics*, 2005.

[24] Liang Dai, Ben Derudder, and Xingjian Liu. Generative network models for simulating urban networks, the case of inter-city transport network in southeast asia. *Cybergeo*, Oct 2016.

[25] Bilal et. al Farooq. Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58:243–263, 2013.

[26] Yafei Han and Jimi Oke. Global urban typology discovery with confirmatory latent class choice model. 2017.

[27] Chauncy D. Harris. A functional classification of cities in the united states. *Geographical Review*, 33(1):86, 1943.

[28] Wagner A Kamakura and Michel Wedel. Exploratory tobit factor analysis for multivariate censored data. 36(1):53–82, 2001.

[29] R. Louf and M. Barthelemy. A typology of street patterns. *Journal of The Royal Society Interface*, 11(101):20140924–20140924, Aug 2014.

[30] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *WALCOM: Algorithms and Computation Lecture Notes in Computer Science*, page 274–285.

[31] ITS Lab MIT. Mobility of the future. https://its.mit.edu/mobility-future.

[32] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295, Oct 2014.

[33] Bengt O. Muthén. Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*, 42(2):241–250, 1989.

[34] D Pellegg and A W Moore. X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[35] Jacob Poushter. Pew research center. http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies.

[36] Roland Priester, Jeffrey Kenworthy, and Gebhard Wulfhorst. The diversity of megacities worldwide: Challenges for the future of mobility. *Megacity Mobility Culture Lecture Notes in Mobility*, page 23–54, 2013.

[37] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[38] Lijun Sun and Alexander Erath. A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61:49–62, 2015.

[39] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[40] Vicapow. Principal component analysis: Explained visually. http://setosa.io/ev/principal-component-analysis/.

[41] J. Sherwood Williams and Dennis Child. The essentials of factor analysis. *Contemporary Sociology*, 3(5):411, 1974.