Showing Connections between Investment and City Growth

Merging statistical and clustering approaches to study growth

Springboard DSI Capstone Project
Final Report by Fernando Meira
July, 2017

ABSTRACT

The debate on measuring growth and comparing different countries and cities is vast. For a long time in sciences, different ways of doing it have been developed. As the times passes, the globalization and the fast-growing world shows that measuring growth is getting more complex each year. This complexity is due to more features that should be taken into consideration and methods that would embrace this multidimensionality in a more natural way. This study is an attempt on tackling this multidimensionality using the I-Distance method and clustering to check the extent to which investment is relevant for city growth and explore the subtleties of development within the province of São Paulo, Brazil.

1. INTRODUCTION

In this section, we are going to cover the foundations of this study. First, we are going to frame the actual debate on how to measure growth and development in countries, which also encompasses the measurement of related cities within the same country. Second, we are going to discuss the main method used for this project, called I-Distance, explaining how it works and its particularities. Finally, we are going to discuss the relevance of this study in the context of Brazil and thereby identifying the specific client for this project, for whom recommendations are suggested at the end of this document.

1.1 The challenges associated with measuring growth

What methodologies to use to measure growth is a topic that has been amply debated in Economics and Social Sciences. Most of the approaches used in this type of studies, during many years, have been focused on just economic aspects. The most used parameter for these has been the GDP (Gross Domestic Product), which is a monetary metric for all the wealth generated in a country. Although this measure can be used to rank countries by their wealth, it has been very poor in capturing socio-economic gaps among the subjects of study.

A first attempt for a ‘multidimensional’ approach started with the creation of the HDI (Human Development Index), by the The World Bank in 1990. This index consists of grouping 3

major indicators: education, income and longevity. Given its simplicity, this index was very popular during a long period but, according to many researchers, it has been criticized because of the large correlation among these indicators and their small number. This oversimplification raised the concern that this index would not capture nuances of the country's development when comparing disparate countries, and in consequence, different scenarios.

As a response to the concerns raised above, some other versions of the index have been created to try to fix these issues, and to try to get more accurate modelling of growth. The most relevant include the following. CDI (Calibrated Human Development Index) which is very similar to HDI but it gives more weight to life expectancy than to education. Another similar is the resource–infrastructure-environment (RIE) index, which attempts to capture structural concerns that are relevant to the characterization of development, including ICT infrastructure variables. Other variables could be added such as internet access, and various similar indicators associated with technology use.

This is an indication that the world nowadays is getting multidimensional really fast and the addition of different features to growth models is going to get more usual than ever. Because of that, approaches that could deal with this multidimensionality in a more natural way, should be of enormous importance within this field of study.

1.2 The I-Distance Method

Considering all the discussion above, the focus for this study was the choice of a multidimensional method that could capture the interaction of different variables of growth, giving the opportunity to check the relative relevance of each of them. The I-Distance method used in this present study was introduced by Milenkovic and other in [1].

The main idea of this method is to use hyperplanes, composed by a given group of chosen variables, to then compare growth between countries (and in our case, cities) by using the distance between different planes. Entities of interest would be represented by the values of chosen variables, through the hyperplane representation, and they would be compared to a unique hyperplane referred to by the authors as the “reference entity”. This reference entity could represent different types of benchmarks that would serve as a baseline for the comparison among the entities. The reference entity could be, for example, the hyperplane computed with the values that can be summarized by using minimum, mean or maximum values.

Given the selected features $X^T = (X_1, X_2, \dots, X_k)$ the I-distance between $e_r^T = (X_{1r}, X_{2r}, \dots, X_{kr}$ and $e_s^T = (X_{1s}, X_{2s}, \dots, X_{ks})$, is defined as follows:

$$D(r,s) = \sum_{i=1}^{k} \frac{|d_i(r,s)|}{\sigma_i} \prod_{j=1}^{i-1} (1 - r_{ji.12\dots j-1})$$

where $d_i(r, s)$ is the distance between the values of variable Xi for e_r and e_s e.g. the discriminate effect,

$d(r,s) = x_{ir} - x_{is}$, $i \in \{1, \ldots, k\}$.

σ_i the standard deviation of X_i, and $r_{ji.12\ldots j-1}$ is a partial coefficient of the correlation between X_i and X_j, ($j < i$).

The I-Distance between the observations and the reference would be given by the discriminate effect divided by the standard deviation of that variable, multiplied by 1 minus the correlation between those variables (which would imply in the “pure”effect of that variable).

In some datasets, negative correlation effects and negative partial correlation effects could occur, mainly in scenarios with reduced number of variables, it would be interesting the use of the I-Distance squared, which is given by:

$$D^2(r,s) = \sum_{i=1}^{k} \frac{d_i^2(r,s)}{\sigma_i} \prod_{j=1}^{i-1} \left(1 - r^2_{ji.12\ldots j-1}\right)$$

There is a second important part of the I-Distance method, which is one of the main points for this project: the possibility of measuring the relevance of each variable for the index calculation.

After the calculation of the I-Distance, the correlation effect of each variable with the I-Distance is calculated. This correlation should be interpreted as the relevance of that variable for the distance index: the bigger the correlation, the most representative that feature is.

Then, after the rank is computed, we choose the smallest correlation value that is not significant. This variable will be excluded from the model, and the calculation (the distance index and the following correlation between features and distance) will be computed again. This process should be repeated until just significant variables remains. At the end, the rank depends only on the variables that are most relevant for measuring growth for that dataset.

1.3 K-Means clustering

The K-Means is in summary a method to classify data into K different clusters, where the value of K is defined *a priori*. K initial centroids are set far as possible from each other and the data points are associated with the nearest centroid.

After this first arrangement, each cluster has its centroids recalibrated to assume the very barycenter of that centroid, and these K centers are relocated and the data points distributed

again. This process is repeated until no more assignments to centroids are possible, and the clustering is given by the last iteration.

When considering the K-Means method in the context of this project, the main approach is to use the clustering method, in addition with the I-Distance ranking values, to look for patterns of growth within the province of São Paulo. This will model different groups of cities that should share some similarities when considering growth as a whole and would allow as to make some comparisons, not just with respect to individual characteristics, also collectively.

1.4 Measuring growth of cities

All the methods discussed above were conceptualized using countries as the entities to be compared. The idea is to consider all of the important aspects related to the development of cities so they can be compared by using a common metric.

Therefore, in our case, once we have all the data that describe the cities under analysis, it is reasonable to use these methods, originally developed to measure countries' growth, to model growth of cities within the same country.

1.5 The relevance of the study and the client

Currently, Brazil seems to be drowning in corruption scandals at various levels. One of the main concerns is the management of public resources and the effect of what is promised and what is built on these promises.

One of the main questions that this study aims to answer is the relevance of investment variables (public and private) in the growth of the cities of the São Paulo province, which might also offer some insights into aspects related to corruption.

The main client for this study is the Office of the Public Governor of the province which could, using the results from this project, measure the relevance of the investment as a policy for growth and balance the distribution of this incentive. We could say that the population of the cities and the province could be an additional stakeholder, who could use the results of this project as an argument for requesting a fairer distribution of resources.

2. THE DATA SET

2.1 Data Presentation

The data set used for this study was merged from two sources:

Investment (public and private) values by year:
http://dados.gov.br/dataset

The rest of the features (demographics, social, economic, etc.):
http://www.imp.seade.gov.br/

The data ranged from the years 2000 to 2016 for all the 645 cities of the province of São Paulo, Brazil. The features of choice were based on the main pillars of growth that are: education, economics, demographics, and social. The complete list of features follows:

Group	Indicator	Source
Demographic Indicators	Population density	IMP Seade - São Paulo
	Eletric energy consumption in Mwh	IMP Seade - São Paulo
	Number of people	IMP Seade - São Paulo
	Number of people living on urban areas	IMP Seade - São Paulo
	Number of people living in rural areas	IMP Seade - São Paulo
	Olding Index (%)	IMP Seade - São Paulo
	Proportion of urban population (%)	IMP Seade - São Paulo
Economic Indicators	GDP per capital (R$)	IMP Seade - São Paulo
	GDP growth (last year - current year) (%)	IMP Seade - São Paulo
	Value Invested by private companies (Million R$)	Dados Gov
	Value Invested by public government (R$)	Dados Gov
	Exportation Values (U$$ FOB)	IMP Seade - São Paulo
	Importation (U$$ FOB)	IMP Seade - São Paulo
	Total Value Added in thousands R$	IMP Seade - São Paulo
	Total Value Added by agriculture in thousands R$	IMP Seade - São Paulo
	Total Value Added by industry in thousands R$	IMP Seade - São Paulo
	Total Value Added by services in thousands R$	IMP Seade - São Paulo
	Total Revenue by Taxes	IMP Seade - São Paulo
	Number of jobs generated	IMP Seade - São Paulo
	Mean jobs income	IMP Seade - São Paulo
Education Indicators	Number of enrolls in undergraduation	IMP Seade - São Paulo
	Number of enrolls in primary school	IMP Seade - São Paulo
	Number of enrolls for the fundamental school	IMP Seade - São Paulo
Social Indicators	Number of policial records by 1000 habitants	IMP Seade - São Paulo
	Human Development Index	IMP Seade - São Paulo
	Total number of hospiral rooms	IMP Seade - São Paulo
	Mean number of hospiral rooms per habitant	IMP Seade - São Paulo

2.2 Data Exploration

The following descriptions and outcomes are available in the Jupiter notebook, available in this capstone project's repository.

2.2.1 Reading the data

The data set was available on a CSV file called SPcities2. All the manipulation and wrangling was done using the Pandas Python package.

2.2.2 Cleaning the data

The main question when exploring the data was to setup the best scenario to apply the I-Distance method. This scenario would be the year with the most data available, so the distance could be more accurate concerning all the features.

The focus for this analysis was on the three main variables for the study: GDP per capita, public and private investment.

When looking to GDP per capita missing values we could see:

```
year
2000      0.0
2001      0.0
2002    645.0
2003    644.0
2004    645.0
2005    645.0
2006    645.0
2007    645.0
2008    645.0
2009    645.0
2010    645.0
2011    645.0
2012    645.0
2013    645.0
2014    645.0
2015      0.0
Name: gdp_per, dtype: float64
```

Figure 1: Number of entries for GDP per capita variable, per year

From this, we conclude that 2000, 2001 and 2015 were years that should not be used in this analysis, since there were no entries for any one of them.

Going a little deeper and trying to see the year with the best combination for the three variables mentioned above, we could see a clear choice as the plot below shows:

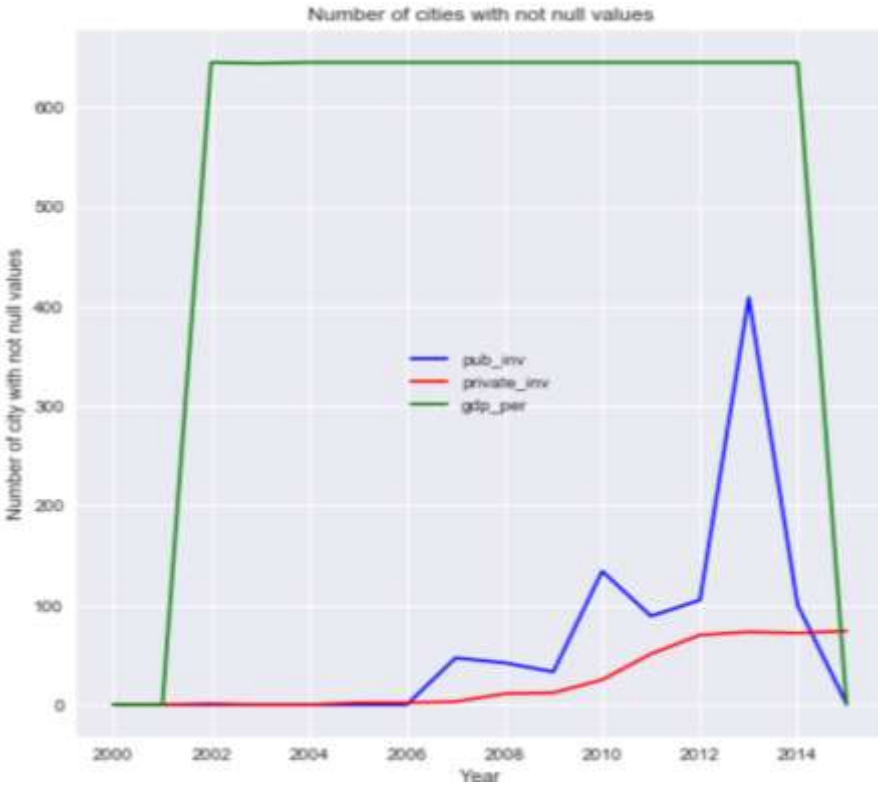

Figure 2: Number of entries for the 3 main features, by year

Looking at this plot the scenario is set. For GDP per capita we can chose from 2001 to 2014, as we see above. But for investment, our best chance is located in 2013, once that private investment is in its top level, and public investment reaches its peak too.

Since year 2013 is a clear choice, our final check would be in the availability of values for the other features in the data-set.

```
city               645
year               645
gdp_per            645
gdp_growth         645
private_inv         73
pub_inv            409
export             354
import             377
violence           645
HDI                  0
educ_superior        0
primary_enrolls    645
density_pop        645
value_add          645
agriculture_add    644
industry_add       645
services_add       645
eletricity         645
tax_revenue          0
hosp_rooms         358
hosp_rooms_per     358
jobs               645
jobs_revenue       645
population         645
urban_pop          645
rural_pop          615
olding             645
urbing             645
fundamental        645
dtype: int64
```

Figure 3: Data entries for all the features in the year of 2013

From the Figure 3, we see that 3 features need to be dropped: HDI, educ_superior and tax_revenue, since none of them have values. For the other variables, we can see that there have plenty of values, which warrants keeping them. Even the private investment that, besides of the fact that it has a small amount of values, it is a crucial variable for the project.

There is an important detail to discuss about missing values. For the computation in the further sections, the missing values were replace with zeros. The main idea here is that, for the most part of the indicators where we don't have all the 645 entries (like public and private investment, hospital rooms, exports and imports, etc.) it is reasonable to assume that, if we have missing values, this occurs because that variable has value equal 0. When looking to different cities is hard to come up with a fair way of 'imputation' since that, if we try assuming the mean for example, we could give this city a characteristic that it should not have.

After this cleaning and dropping the 3 features above, the data set was ready for further analysis.

3. DATA ANALYSIS AND RESULTS

3.1 Defining the Reference Entity

As defined above, the reference entity is the foundation for the calculation of the I-Distance index. Among the many possibilities for this element, in this project the focus was entirely on

the indicator GDP per capita. This was the case because the GDP per capita, besides of the fact that is a very used parameter in another growth studies, it resumes a decent part of wealth distribution because of its 'normalization' by the number of habitants.

Another important fact was that, for this project, the reference entity was chosen to be a real city, which could assume a more real point of comparison than taking the mean or median of all indicators. Following this thought, the city of Itapirapuã Paulista was chosen because it has the smallest number for GDP per capita among all the cities in the studied period. This implies that cities with great values of I-Distance will reflect, by the definition of this metric, cities far from the reference and as consequence, cities with more growth.

As result, the reference entity vector would assume this form:

Indicator	Value
city	Itapirapuã Paulista
year	2013
gdp_per	6956,9
gdp_growth	0,098175775
private_inv	0*
pub_inv	917044,27
export	0*
import	0*
violence	12
HDI	0*
educ_superior	0*
primary_enrolls	102,25
density_pop	9,73
value_add	26929,78
agriculture_add	3766,22
industry_add	1356,65
services_add	21806,9
eletricity	2302
tax_revenue	0*
hosp_rooms	0*
hosp_rooms_per	0*
jobs	521
jobs_revenue	1266,38
population	3954
urban_pop	1959
rural_pop	1995
olding	42,32
urbing	49,54
fundamental	64

** These values were missing values at first, but they were filled with 0 as discussed in section 2.2.2*

Table 1: Values of the reference for each indicator

3.2 Choosing between two versions of the I-Distance method

As defined above, the I-Distance method was the main tool for this analysis. Recall there are two ways of using the I-Distance method: the squared version, and the not squared one.

The not squared version is clearly more sensitive, to big differences in variable values and small number of data (or features), which could affect the final results. To check for the difference in the results, the two versions were applied and the 'sensitivity' of the non-squared method is clear as seen in the table below:

Method	Number of iterations	Number of excluded features	Features that remained
Not Squared	20	20	gdp_growth; gdp_per; density_pop; urbing
Squared	7	7	violence; jobs; services_add; valued_add; urban_pop; population; hosp_rooms; fundamental; pub_inv; electricity; industry_add; rural_pop; import; export; density_pop; jobs_revenue; gdp_per

Table 2: Number of iterations and excluded features for the I-Distance calculation

As we can see, the non-squared version eliminated a big part of the features that, in the squared version, remained. This is taken as a negative point for this version of the model, since the main point of the study is to check for the features that contribute the most to growth and how they contribute to it. In addition to this, for the rank of the features of the non-squared version, we can see:

	feature	p-value	r
1	gdp_growth	2.519221e-61	0.588352
2	gdp_per	1.639504e-08	0.220019
3	density_pop	3.476726e-02	-0.083140
4	urbing	7.199990e-05	-0.155650

Figure 4: Rank of features relevance for the not squared I-Distance

Since the 'r' column reflects the correlation of that feature with the I-Distance calculation (as defined in the introductory section), the contribution of that variable, for "density_pop" and "urbing" features we have a negative 'r', which makes the interpretation difficult, since these variables are still significant. It is important to say here that, the p-value in this context

represents the probability of finding the value on the table above if in fact the value of ‘r’ were 0.

Then we can conclude here that we should focus more on the squared version.

3.3 The results of the squared I-Distance method

As mentioned before, for this version of the method, 7 iterations were needed and the following variables were excluded as non-significant, in this order: olding, agriculture_add, gdp_growth, hosp_rooms_per, primary_enrolls, private_inv and urbing.

Taking the first look at the features, we can see the rank of relevance as follows:

	feature	p-value	r
1	violence	0.000000e+00	0.990580
2	jobs	0.000000e+00	0.989353
3	services_add	0.000000e+00	0.985795
4	value_add	0.000000e+00	0.982313
5	urban_pop	0.000000e+00	0.978691
6	population	0.000000e+00	0.978607
7	hosp_rooms	0.000000e+00	0.976214
8	fundamental	0.000000e+00	0.972127
9	pub_inv	0.000000e+00	0.951517
10	eletricity	8.198462e-314	0.944806
11	industry_add	4.002641e-288	0.933264
12	rural_pop	4.962174e-113	0.740399
13	import	1.767463e-86	0.673588
14	export	3.427779e-85	0.669827
15	density_pop	8.103492e-13	0.276907
16	jobs_revenue	3.553167e-05	0.162043
17	gdp_per	1.193716e-02	0.098937

Figure 5: Rank of features relevance for the squared I-Distance

One of the most interesting points that we are seeking in this project is to check the relevance of investment variables in the calculations. As we can see, the private investment was dropped in the sixth iteration. This implies that, for this scenario, the private investment was not important for the growth of the cities. On the other hand, the public investment figured as a relevant feature, located in the middle of the rank. Another interesting result concerning the features, is that, although we have a lot of economic features in the top of the rank (jobs, service_add, value_add) the top feature was the violence indicator. This is a very

interesting finding, given that violence is one of the main weak points pertaining public management in many cities of all the country.

For the rank of cities, we have:

	I-Distance	city
1	5420.100823	São Paulo
2	253.370130	São Sebastião
3	150.584562	São José dos Campos
4	142.742285	Ilha Comprida
5	120.551111	São Bernardo do Campo
6	114.357130	Diadema
7	112.603475	Santos
8	112.293461	Louveira
9	104.851029	Taboão da Serra
10	102.321066	Barueri
11	100.644117	Osasco
12	91.586236	São Caetano do Sul
13	87.955723	Guarulhos
14	85.668883	Campinas
15	78.343459	Carapicuíba
16	58.143407	Paulínia
17	43.994971	Alumínio
18	41.225816	Jundiaí
19	40.666734	Sorocaba
20	39.531636	Mauá
21	36.173007	Santo André
22	35.980659	Jaguariúna
23	34.704392	Piracicaba
24	33.638510	Cajamar
25	32.175117	Cubatão

Figure 6: Rank of cities for the squared I-Distance (best positioned cities)

For the cities, we can see that a very good reality check for the index is that the city of São Paulo is on the top of the rank. São Paulo is a very big city, with hosting many of global companies, it is a crucial point for the financial industry of the country, and it is clearly an outlier in the province. Other cities that are natural candidates to be placed in the best positions of the rank: Campinas, São José dos Campos, Santos, Guarulhos and Osasco. These cities are relatively big, well positioned in the logistics scheme for industries and core technology development centers within the province and even within the country. Other cities of the province that are not that relevant when looking to common parameters like São Sebastião, Ilha Comprida and Louveira, are big surprises here and should be a very

interesting start point when looking through the implications of their position and the parameters they have.

3.4 Merging the I-Distance method with clustering

When looking for growth contributors and ranking cities, one of the most common questions is if there are patterns or related groups of cities that would make sense geographically, economically or even demographically.

To tackle this point, we used clustering methods (K-means) to combine the index explored above and look for patterns within the province.

The number of centroids (K) was chosen using the so-called Elbow Method. The number chosen was 5, and this choice is clear when looking at the following plot:

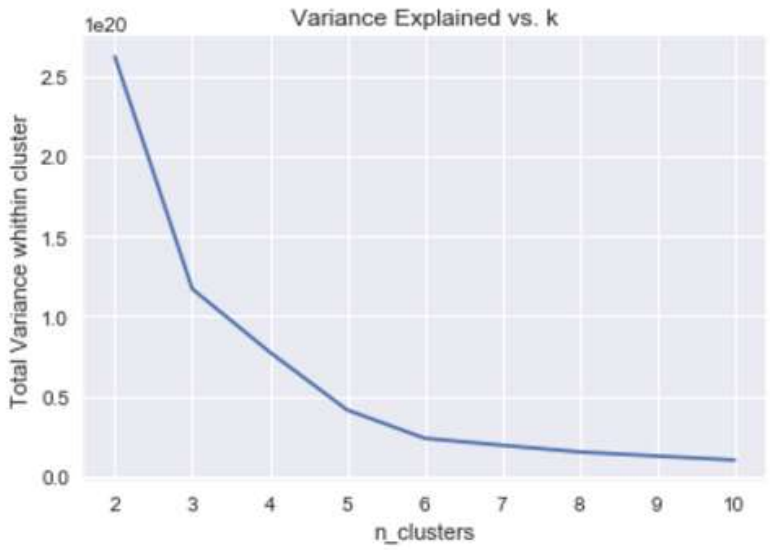

Figure 7: Variance explained versus the number of centroids chosen

With the optimal number of centroids chosen, the clusters are shown in the following figure:

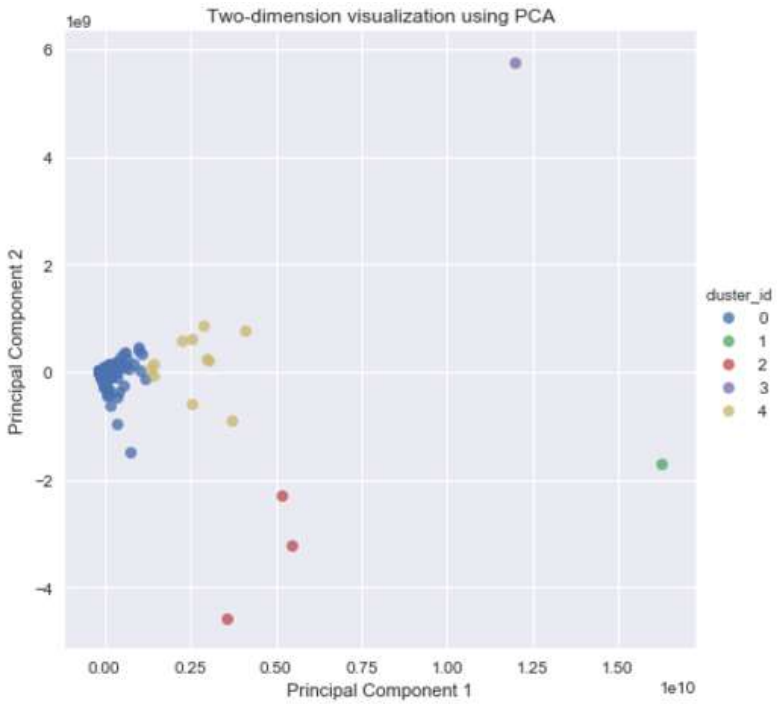

Figure 8: The distribution of the clusters using all the features

For the visualization of the Figure 8 to be possible in a two-dimensional plot, a Principal Component Analysis (PCA) was performed. Briefly, PCA is an algorithm used to reduce the dimensionality of a data set. The main idea is to reduce the number of dimensions using variables that are correlated (heavily or lightly), transforming the correlated variables in new set of variables (called principal components) and preserving the variation of the data as maximum as possible. For the Figure 8, a PCA, consisting of 2 principal components, was computed, reducing the dimension of the dataset to 2 and making possible the two-dimension visualization of the data. Also, the cities are colored according to the clusters they belong to.

Looking to the figure above it's clear why the number of 5 cluster is the optimal solution. There are clearly 5 different groups and the calculations for most representative cities for the centroids shows:

Cluster_id	Most representative city	I-Distance Squared Rank
0	Porto Feliz	228
1	São Paulo	1
2	Taubaté	29
3	São Sebastião	2
4	São José dos Campos	3

Table 3: Cluster most representative cities

The ‘one-city’ clusters are represented by São Paulo (for cluster 1) and São Sebastião (cluster 3). This seems to be natural for the case of the city of São Paulo, that is a clear outlier. When looking to the case of the cluster represented by the city of São Sebastião, this is not so clear, considering that it is a medium city, located in the coast of the province and far from the port infrastructure that is in the south part of the coast. However, these two cities are ranked as the best two cities when considering the index, even if a big gap separates them. This implies that, despite the territory size or infrastructure, the complete set of feature should be considered, and furthermore, this single fact helps corroborating the assumption that growth must take into account multiple variables.

	Cluster	I-Distance	city
4	0	142.742285	Ilha Comprida
6	0	114.357130	Diadema
8	0	112.293461	Louveira
9	0	104.851029	Taboão da Serra
11	0	100.644117	Osasco
12	0	91.586236	São Caetano do Sul
15	0	78.343459	Carapicuíba
17	0	43.994971	Alumínio
20	0	39.531636	Mauá
22	0	35.980659	Jaguariúna
24	0	33.638510	Cajamar
228	0	1.809157	Porto Feliz

Figure 9: Rank of cities for the squared I-Distance (best positioned cities) for cluster 0 with the cluster centroid highlighted

	Cluster	I-Distance	city
1	1	5420.100823	São Paulo

Figure 10: Rank of cities for the squared I-Distance (best positioned cities) for cluster 1 with the cluster centroid highlighted

	Cluster	I-Distance	city
3	2	150.584562	São José dos Campos
5	2	120.551111	São Bernardo do Campo
7	2	112.603475	Santos

Figure 11: Rank of cities for the squared I-Distance (best positioned cities) for cluster 2 with the cluster centroid highlighted

	Cluster	I-Distance	city
2	3	253.37013	São Sebastião

Figure 12: Rank of cities for the squared I-Distance (best positioned cities) for cluster 3 with the cluster centroid highlighted

	Cluster	I-Distance	city
10	4	102.321066	Barueri
13	4	87.955723	Guarulhos
14	4	85.668883	Campinas
16	4	58.143407	Paulínia
18	4	41.225816	Jundiaí
19	4	40.666734	Sorocaba
21	4	36.173007	Santo André
23	4	34.704392	Piracicaba
29	4	26.810150	Taubaté
36	4	19.950008	Sumaré
41	4	15.174380	Indaiatuba

Figure 13: Rank of cities for the squared I-Distance (best positioned cities) for cluster 4 with the cluster centroid highlighted

On the other hand, when considering the question of whether investment (in this case public) is relevant for the growth of these cities, we can draw very interesting insights by grouping the cities with the cluster labels and comparing these as different groups.

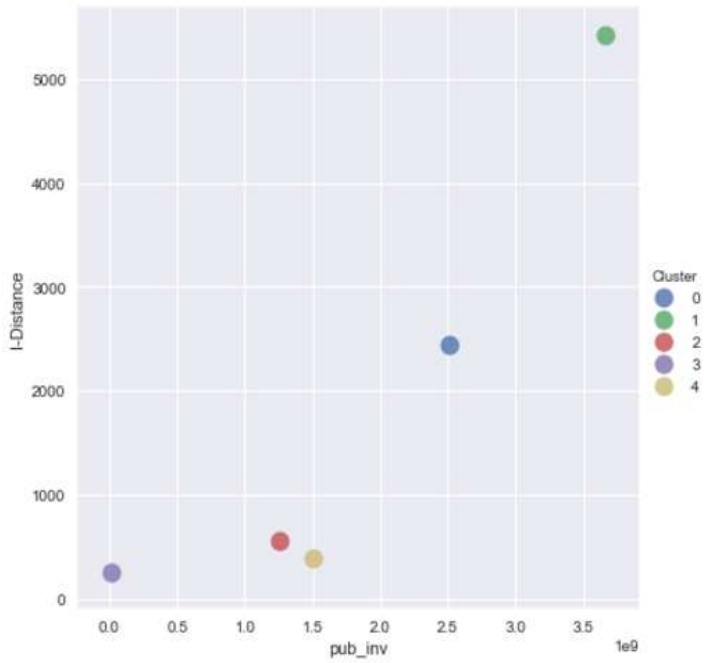

Figure 10: Scatter plot of the sum of I-Distance and public investment by cluster

Although cluster 0 groups the largest number of cities, there is a clear visual relationship where, the clusters with more absolute value of investment have better index. However, we

should admit that, when using the mean or median (as shown in Figure 10), to normalize this distribution effect, the relationship is way subtler.

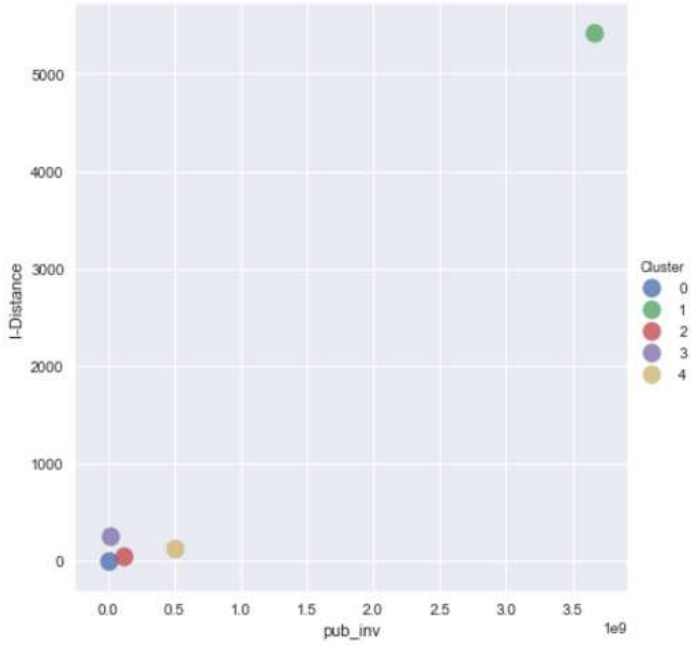

Figure 11: Scatter plot of the mean of I-Distance and public investment by cluster

4. CONCLUSIONS

This project explored the extent on which investment is relevant for city growth, and as consequence, how growth differs from city to city (or group of cities).

Using the I-Distance as a unifying metric, it seems that private investment was not relevant in any of the scenarios we explored. However, public investment is shown to be relatively important. Moreover, when looking at the clusters computed using the K-Means algorithm, public investment seems to be an important variable that can be used to differentiate cities according to their growth, when considering different types of cities within the province (according to the features used in this project.).

The models we built using the concept of I-Distance, and its combination with clustering were relatively good, positioning cities with theoretical growth in expected places and also showing cities that are not on the radar of many public agents, but should be studied at a deeper level.

5. FUTURE WORK

Possible points for further development include the following.

- Take a deeper look at features other than investment that are relevant to the computed clusters.
- A deeper study of the high positioned cities that were not expected and check which attributes made them go higher than others.
- Add more features related to technology, such as internet access.
- Look at data for more recent years that should have more entries for the private investment feature, to confirm if it is not relevant as found in this study.

6. RECOMMENDATIONS FOR THE CLIENT

As part of the recommendations for the client, we would like to highlight the following:

- Better management of public investment funding, since it seems to have relative relevance for cities growth.
- Look for benchmarks outside well-known cities, considering different features in the analysis.
- Take a closer look at violence indicators, since it was classified as one of the most important features on the subset we studied.

6. CONSULTED REFERENCES

---

[1] N. Milenkovic, et alia: "A multivariate approach in measuring socio-economic development of MENA countries". In Economic Modeling 38 (2014) 604-608. Available from **http://www.sciencedirect.com/science/journal/02649993/38**

[2] J. Bang, et alia: "New Tools for Predicting Economic Growth Using Machine Learning: A Guide for Theory and Policy".. Available from: **https://www.researchgate.net/publication/291827961_New_Tools_for_Predicting_Economic_Growth_Using_Machine_Learning_A_Guide_for_Theory_and_Policy**

[3] N. Adriansson, et alia: "Forecasting GDP Growth, or How Can Random Forests Improve Predictions in Economics? " . Available from: **https://pdfs.semanticscholar.org/d402/473ba628b67bcd0d3a8cf39799ae6efbdc66.pdf**