Is Investment Important for Cities Growth?

Merging statistical and clustering approaches to study growth

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 were developed. As the times passes, the globalization and the fast growing world shows that measuring growth it`s 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 the multidimensionality using the I-Distance method and clustering to check if investment is relevant for cities growth and explore the subtleties of the 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 measuring growth and development in countries (and for consequence, the same method for cities). Second, we are going to introduce the main method used for this study called I-Distance, explaining the method workflow and particularities. And the third and last item for this section, will be the discussion on the relevance of this study in the present scenario of Brazil and the definition of the main client for the results to be found in the end.

1.1 The challenges on measuring growth

The methodology on measuring growth is a very debated topic in Economics and Social Sciences. The majority of the approaches used in this type of studies, during many years, were focused just on the economic side. The most used parameter for these was the GDP (Gross Domestic Product), a monetary account for all the wealth generated in a country. In spite of the fact that this measure could rank countries by their wealth, it was very poor on capturing socio-economic gaps within the objects 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 on grouping 3 major indicators: education, income and longevity. Given its simplicity, this index was very famous during a long period but, by many researchers, it was criticized by the big correlation among the features and the small number of variables.This oversimplification raised the concern that this index would not capture nuances of the country's development when looking to different countries, and as consequence, different scenarios.

As a response to the questions raised above, some different versions of the index were created trying to fix this issues and getting more accurate perceptions of growth. Some most relevants are: 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 includes more structural concerns to the development debate, including ICT infrastructure variables. Another many different ways on trying to approach growth as a multidimensional question includes also the addition of features like internet access and similar derivations for technology use.

This only shows 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 study field.

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 on growth, giving the opportunity to check the relative relevance of each of them. The I-Distance method used in this present study is most inspired by the definition of Milenkovic, 2014 [1] .

The main idea of this method is to use hyperplanes, composed by a given group of chosen variables, to compare growth between countries (and in our case ,cities) using the distance between different planes. The observations would constitute the interest entity and they would be compared to a unique hyperplane called by the author above as the "reference entity". This reference could assume different types of benchmarks that would serve as a baseline for the study. It could be, for example, the observation with a minimum, mean or maximum value for some variable (or a group of them, if possible) .

Given the features selected $X^T = (X_1, X_2, \dots , X_k)$ and the separate hyperplanes given by $e_r = (X_{1r}, X_{2r}, \dots , X_{kr})$, e.g reference, and $e_s = (X_{1s}, X_{2s}, \dots , X_{ks})$, e.g the observations, the equation for the I-distance would be defined as:

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

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,

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

σ_i the standard deviation of X_i, and $r_{ji\cdot 12 \dots j-1}$ 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 sets, 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:

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

There is a second important part of the I-Distance method, which is one of the main focus for this study: the possibility of measure 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 done, we choose the smallest correlation value that is not significant (that should be represented by $p - value > \alpha$, at the chosen threshold α). This variable will be excluded from the model, and the calculation (the distance index and the following correlation between features and distance) will be done again. This process should be repeated until just significant variables remains. By the end, a rank of the most relevant features should remain, indicating only the variables that are relevant for measuring growth for that set.

1.3 K-Means clustering

The K-Means is in summary a method to classify data into k different cluster defined *a priori*. K initial centroids are set far as possible from each other and the data points are distributed to each nearest centroid.

After this first arrange, each cluster has its centroids recalibrated to assume the very barycenter of that centroid, and this k centers are relocated and the data points distributed again. This process is done until no more move by the centroids are done, and the clustering is given by last iteration.

1.4 Measuring cities growth

All of the methods discussed above were idealized around country's growth studies. The concern in considering all of the important aspects of the development of nations has, in its foundation, the question on how different people, with different government systems, healthcare programs, economies, territory sizes, level of violence, etc, grow the wealth of their nation.

This elementary question is possible because, as we know, each country has it's own particularities, and it`s own data that describe its situation. This is very similar when we move to cities discussion. It's is very common, mainly in medium to big countries (like the case of this study - Brazil), to select two cities that belongs to the same province but have very different situation on culture, economic growth, healthcare, etc.

So, in our case, once we have all the data that describes these particularities, is reasonable to extend this methods used for countries growth, to test data on cities too.

1.5 The relevance of the study and the client

By the moment this study is done, all brazilians are drowned in corruption scandals from politics that develop in many outcomes that range from gossips to economic crisis. One of the main concerns is the management of public resources and the effect of what is promised and what is built with the promises.

One of the main questions that this study aims to answer is the relevance for the investment variables (public and private) in the growth of the cities of the São Paulo province, which is very related to the corruption chapters.

The main client for this study is the Public Governor of the province which could, with the results, 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 a additional client, which could use the results as an argument for claiming for a more fair management.

2. DATA SET

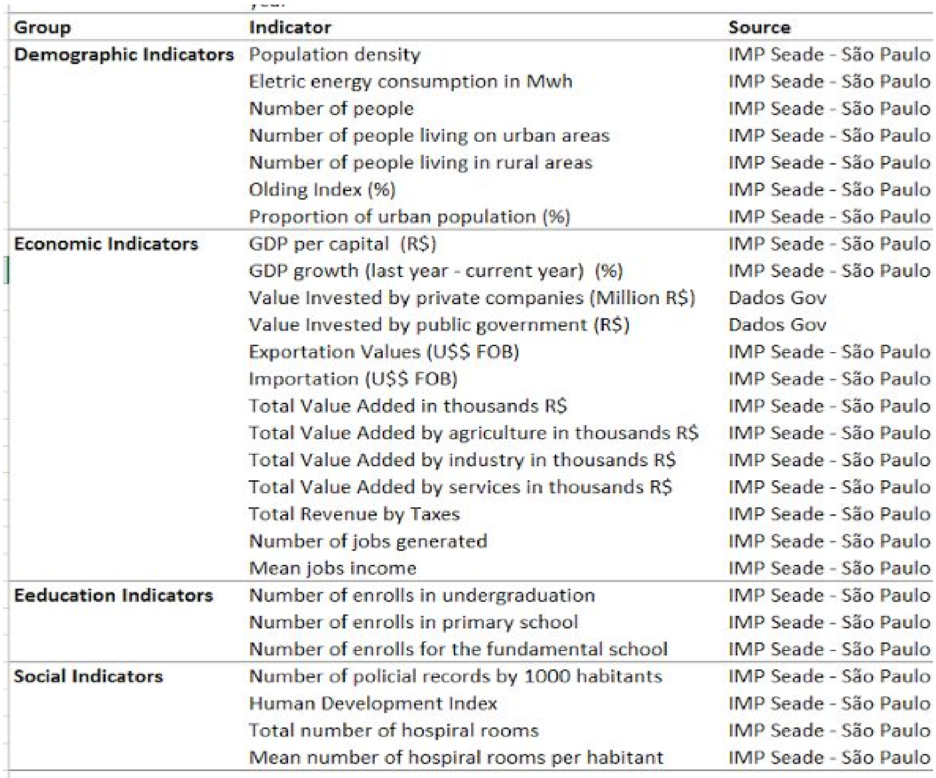
2.1 Data Presentation

The data set used for this study was a merge of 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 of 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 list of features are:



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
Eeducation 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 on the ipynotebook file, on the capstone repository.

2.2.1 Reading the data

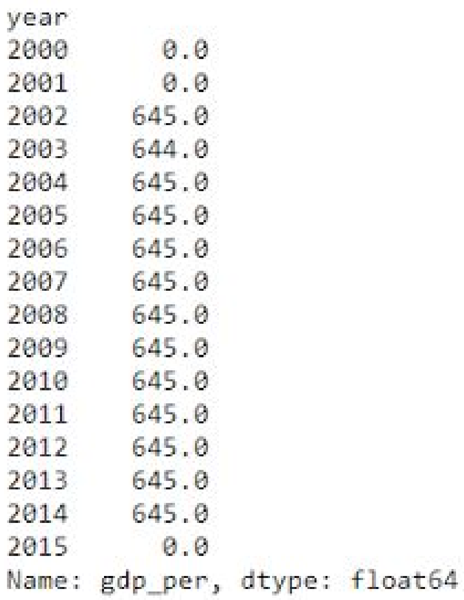
The data set was available on a csv file called SPcities2.csv. All the manipulation and wrangling was done using Pandas.

2.2.1 Cleaning the data

The main question when looking the data was to set 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

Here we could conclude that 2000, 2001 and 2015 were years that should not be used to this analysis, once 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:

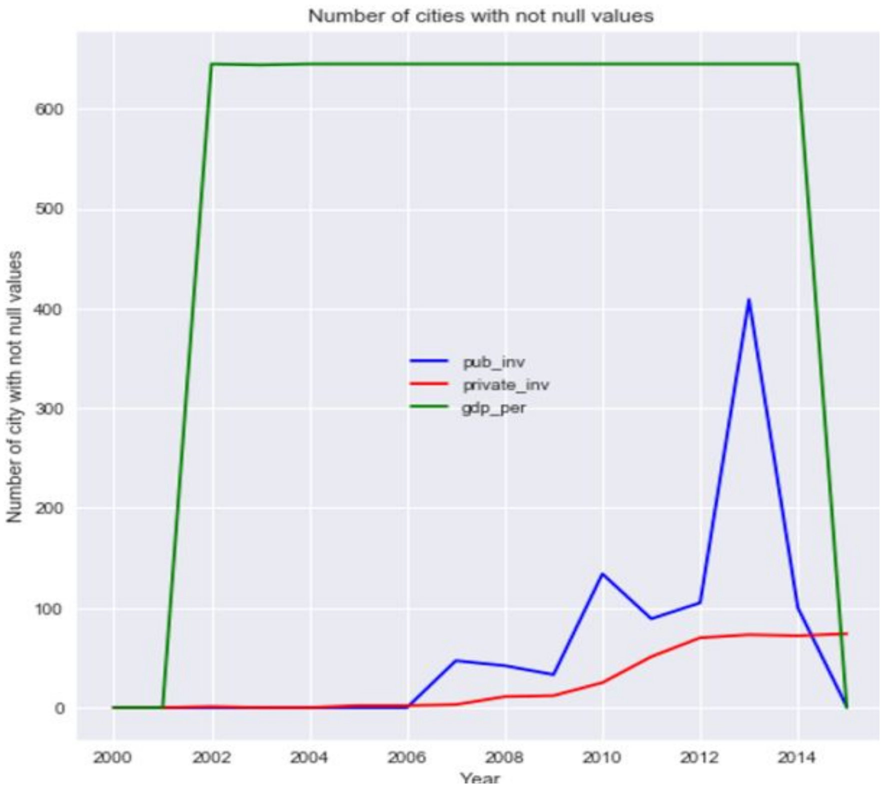
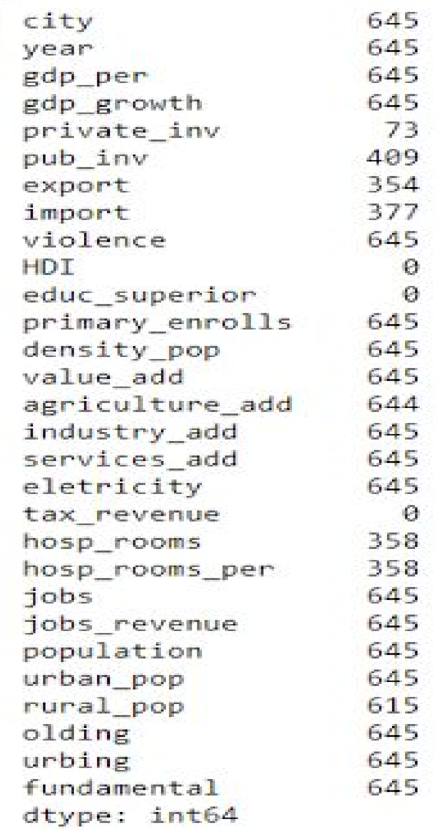


Figure 2: Number of entries for the 3 main features, by year

Looking at this plot the scenario is set. For GDP per capita we can chose from 2001 to 2014, as we seen 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 peek too.

Once that the year of 2013 is a clear choice, our final check would be in the values availability 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

When looking to the Figure 3 we could see that 3 features need to be dropped: HDI, educ_superior and tax_revenue.

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

3. RESULTS

3.1 Choosing between two versions of the I-Distance method

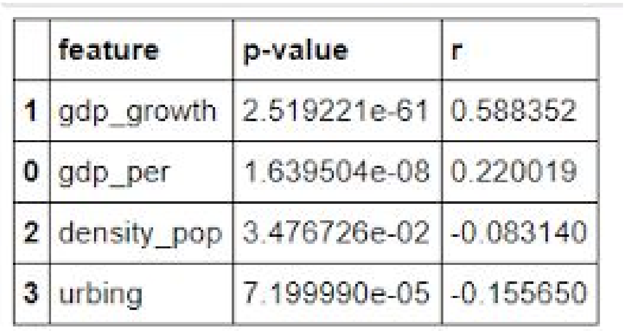
As defined above, the I-Distance method was the main tool for this analysis. One of the first things we need to explain here is that, the 2 ways of using the method were applied: the squared version and the not squared.

The not squared version is clearly more sensible once that, big differences in variable values and small number of data (or features), could affect the results. To check for the difference in the results, the two versions were applied and the 'sensibility' of the not squared 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 1: Number of iterations and excluded features for the I-Distance calculation

As we can see, the not squared version eliminated big part of the features that, in the squared version, remained. This is taken as a negative point for this model, once that the main interest of the study is to check for the features that contribute more to growth and how they contribute. Beside of this fact, for the rank of the features of the not squared version, we can see:



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

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

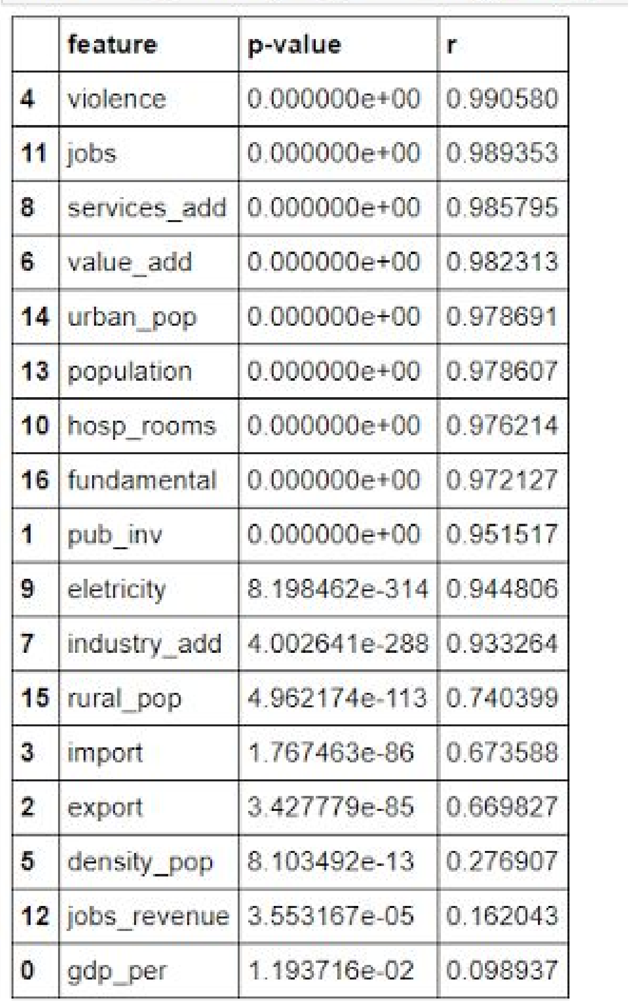
Once that the ‘r’ column reflects the correlation of that feature with the I-Distance calculation (as defined in the introduction section), hence the contribution of that variable, for density and urbing feature we have a negative ‘r’, which makes the interpretation difficult, once that these variables are still significant.

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

3.2 The results of the squared I-Distance method

As said above, 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:

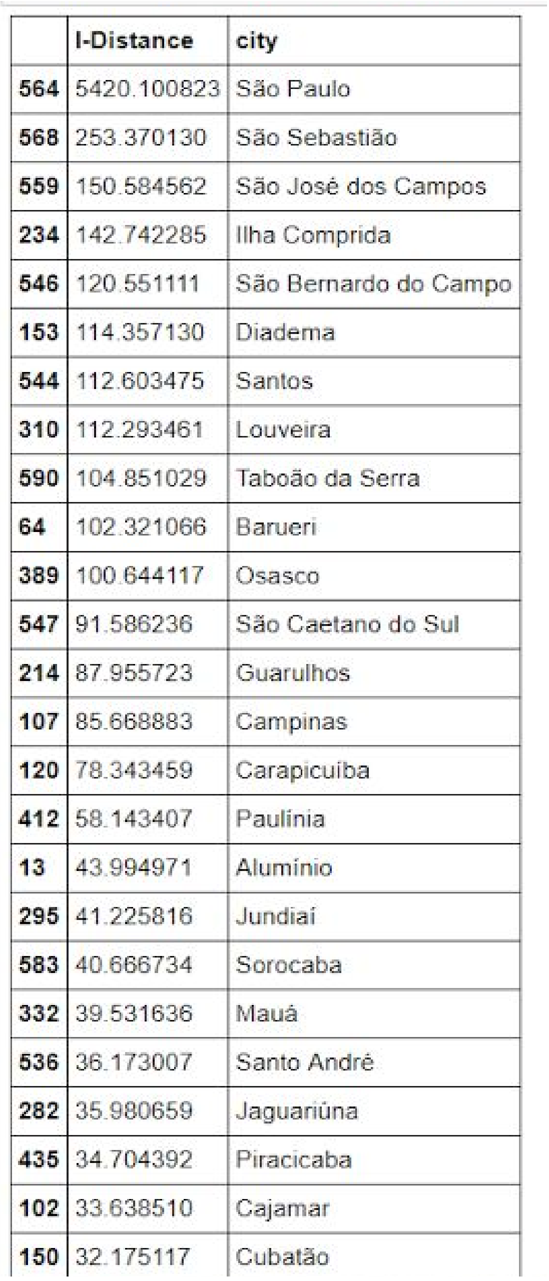


	feature	p-value	r
4	violence	0.000000e+00	0.990580
11	jobs	0.000000e+00	0.989353
8	services_add	0.000000e+00	0.985795
6	value_add	0.000000e+00	0.982313
14	urban_pop	0.000000e+00	0.978691
13	population	0.000000e+00	0.978607
10	hosp_rooms	0.000000e+00	0.976214
16	fundamental	0.000000e+00	0.972127
1	pub_inv	0.000000e+00	0.951517
9	eletricity	8.198462e-314	0.944806
7	industry_add	4.002641e-288	0.933264
15	rural_pop	4.962174e-113	0.740399
3	import	1.767463e-86	0.673588
2	export	3.427779e-85	0.669827
5	density_pop	8.103492e-13	0.276907
12	jobs_revenue	3.553167e-05	0.162043
0	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 were seeking in this study was 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, in spite of the fact that 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 founding, once that violence is one of the main weak points of the public management in many cities of all the country.

For the rank of cities we have:



	I-Distance	city
564	5420.100823	São Paulo
568	253.370130	São Sebastião
559	150.584562	São José dos Campos
234	142.742285	Ilha Comprida
546	120.551111	São Bernardo do Campo
153	114.357130	Diadema
544	112.603475	Santos
310	112.293461	Louveira
590	104.851029	Taboão da Serra
64	102.321066	Barueri
389	100.644117	Osasco
547	91.586236	São Caetano do Sul
214	87.955723	Guarulhos
107	85.668883	Campinas
120	78.343459	Carapicuíba
412	58.143407	Paulínia
13	43.994971	Alumínio
295	41.225816	Jundiaí
583	40.666734	Sorocaba
332	39.531636	Mauá
536	36.173007	Santo André
282	35.980659	Jaguariúna
435	34.704392	Piracicaba
102	33.638510	Cajamar
150	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, it has a lot of global companies installed, it is a crucial point of the financial industry of the country, and it is clearly an outlier in the province and in the country. Other cities that are natural candidates for the top of the rank are there like: Campinas, São José dos Campos, Santos, Guarulhos and Osasco. These cities are relative 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.

3.2 Merging the I-Distance method with clustering

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

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

Before the results, we should explicit here that the number of centroids (k) was chosen using the Elbow method. The number chosen was 5, and this choice is clear when looking the following plot:

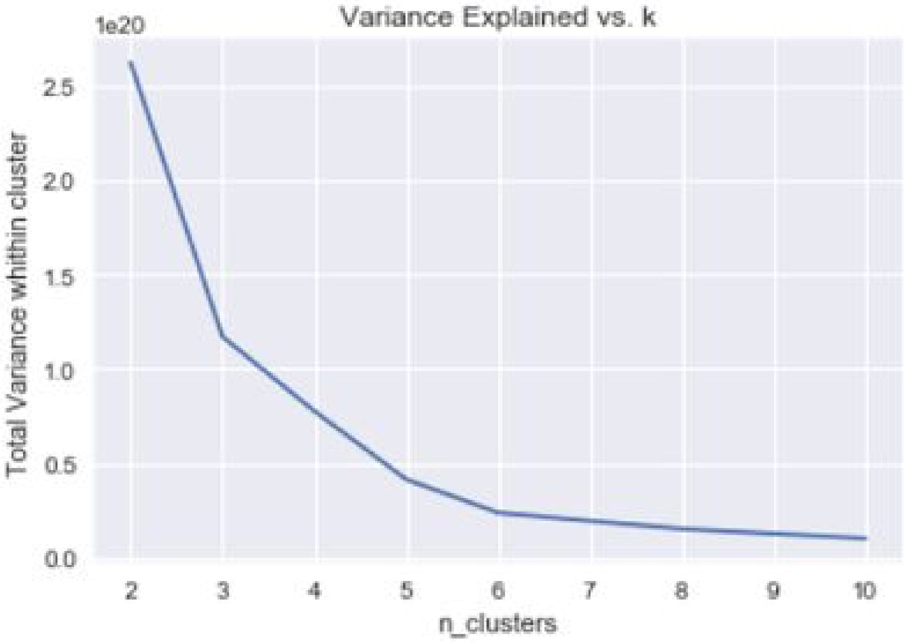


Figure 7: Variance explained versus the number of centroids chosen

With the optimal number of centroids chosen, the clusters could be seen as the following figure:

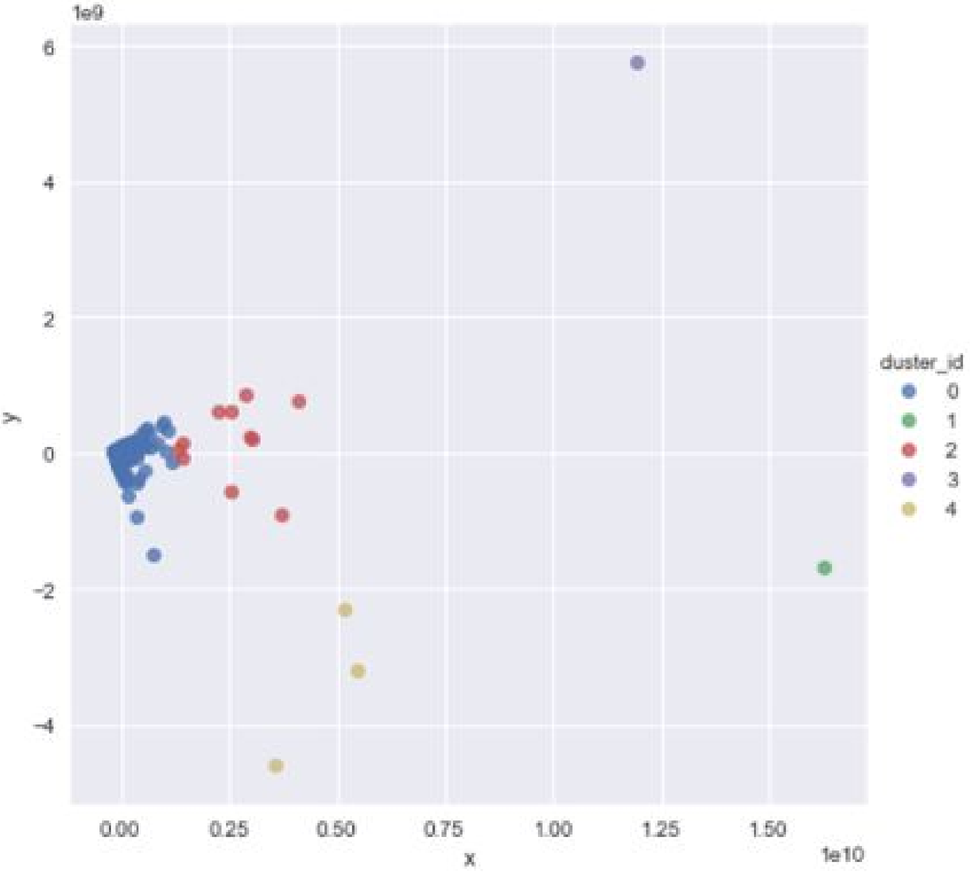


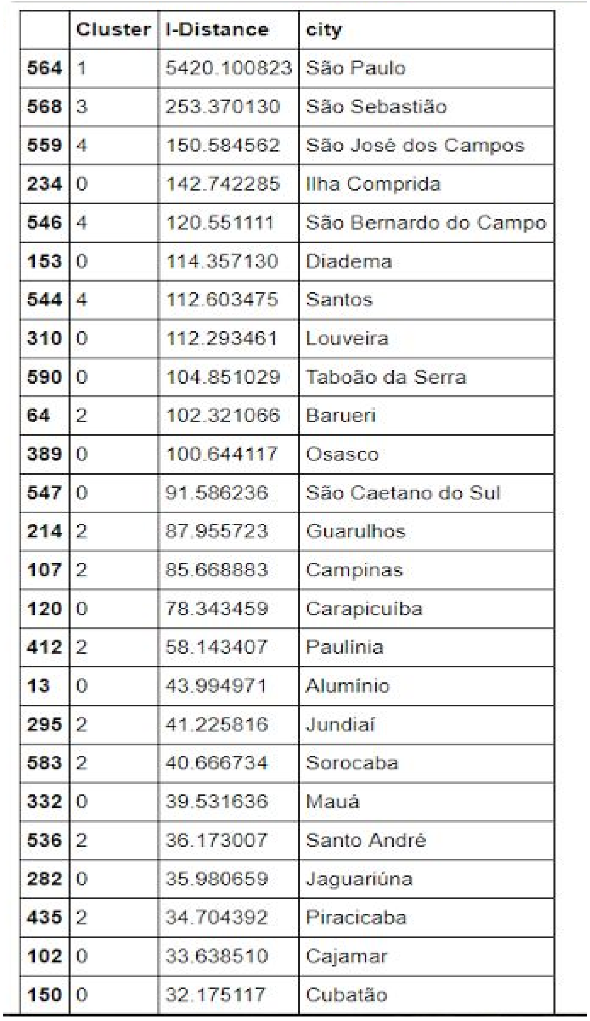
Figure 8: The distribution of the clusters using all the features

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

Cluster_id	Most representative city
0	Porto Feliz
1	São Paulo
2	Taubaté
3	São Sebastião
4	São José dos Campos

Table 2: 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 located in the south part of the coast. However, these two cities are ranked as the best two cities when considering the index. This implies that, in spite of the territory size or infrastructure, the complete set of feature should be considered, and even more, this single fact helps corroborating the assumption that growth is a multidimensional measure.



	Cluster	I-Distance	city
564	1	5420.100823	São Paulo
568	3	253.370130	São Sebastião
559	4	150.584562	São José dos Campos
234	0	142.742285	Ilha Comprida
546	4	120.551111	São Bernardo do Campo
153	0	114.357130	Diadema
544	4	112.603475	Santos
310	0	112.293461	Louveira
590	0	104.851029	Taboão da Serra
64	2	102.321066	Barueri
389	0	100.644117	Osasco
547	0	91.586236	São Caetano do Sul
214	2	87.955723	Guarulhos
107	2	85.668883	Campinas
120	0	78.343459	Carapicuíba
412	2	58.143407	Paulínia
13	0	43.994971	Alumínio
295	2	41.225816	Jundiaí
583	2	40.666734	Sorocaba
332	0	39.531636	Mauá
536	2	36.173007	Santo André
282	0	35.980659	Jaguariúna
435	2	34.704392	Piracicaba
102	0	33.638510	Cajamar
150	0	32.175117	Cubatão

Figure 9: Rank of cities for the squared I-Distance (best positioned cities) with clusters

On the other hand, when looking to the question on whether investment (in this case the public, that was the only that remained in the model) is relevant to the growth of this cities, we can have very interesting insights grouping the cities with the cluster labels and comparing these as different groups.

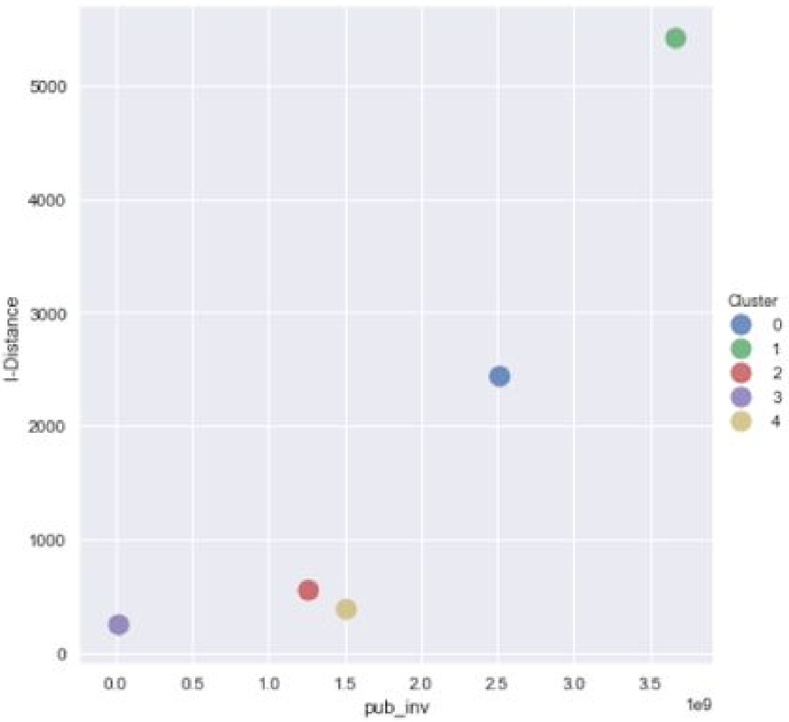


Figure 10: Scatter plot of the sum of I-Distance and public investment by cluster

In spite of the fact that the cluster 0 has the most part of the 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 more subtle.

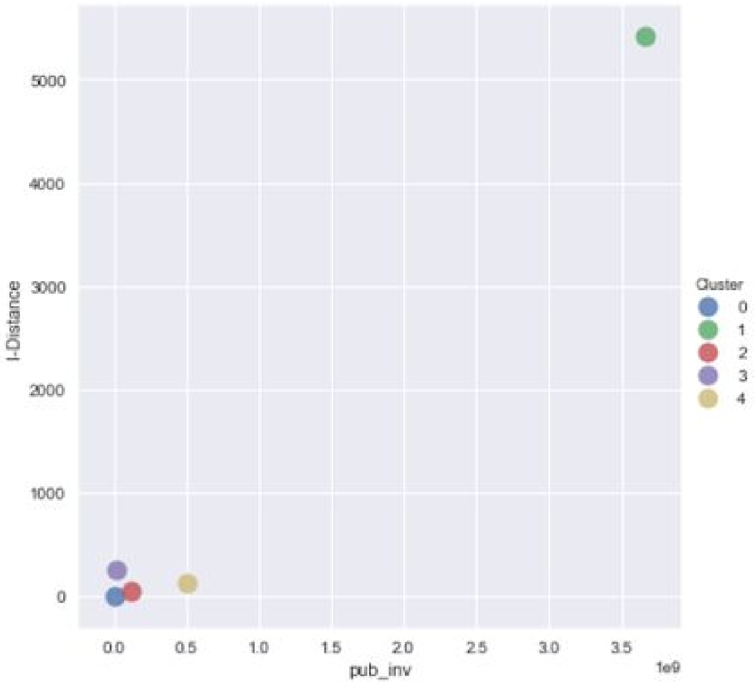


Figure 11: Scatter plot of the mean of I-Distance and public investment by cluster

4. CONCLUSION

As the results seen above, this study aimed the question on whether investment are relevant to city growth, and as consequence, how growth differs from city to city (or group of cities).

We could conclude that private investment was not relevant in any of the scenarios. On the other hand, the public investment was relative important, considered a medium relevant feature. In addition, when looking into the clusters, public investment seems to be one good approach to differentiate city growth, when considering different types of cities within the province (considering the features of this study, of course).

As a final conclusion, our model was relatively good, positioning cities with theoretical growth in expected places and showing cities that are not on the radar of many public agents, but should be studied in a deeper level.

5. FUTURE WORK

Thinking on further analysis, some important points would be:

- Take a deeper understanding on other features than investment that are relevant to the clusters found.
- 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 like internet access.
- Look to 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 recommendations for the client, we should highlight:

- Better management on public investment funding, once it is has relative relevance for cities growth.
- Look for benchmarks outside the most known cities, considering different features in the analysis.
- A more delicate action on the violence indicators, once that it was classified as the most important feature on the subset 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

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

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