




Predicting innovative cities using spatio-temporal activity patterns

Ricardo Muñoz-Cancino¹, Sebastián A. Ríos², and Manuel Graña¹

Computational Intelligence Group, University of Basque Country, 20018 San Sebastián, Spain.

² Business Intelligence Research Center (CEINE), Industrial Engineering Department, University of Chile, Beauchef 851, Santiago 8370456, Chile



Abstract

Understanding cities' complexity is essential for correctly developing public policies and urban management.

We attempted to relate the activity carried out by city inhabitants with the macro characteristics of a city, mainly its capacity to innovate.

In this study, we seek to find those features that allow us to distinguish between an innovative city from those still on the way to becoming one.

The innovation index is a characteristic measuring the capacity and development of cultural assets, infrastructure, and the quality of markets.

To carry out this analysis, we have the activity patterns decomposition obtained through geo-tagged social media digital traces and their respective innovation index for more than 100 cities worldwide.

The results show that it is possible to predict the city's innovative category from their activity patterns.

Our model achieves an $AUC = 0.71$ and a $KS = 0.42$. This result allows us to establish a relationship between the activities carried out by people in the city and its innovation index,

Contents

- Introduction
- Dataset description
- Target variable
- Data augmentation
- Overall methodology
- Experimental design
- Results
- Conclusions

Introduction

Innovation is critical to address challenges such as urbanization, social equality, and climate change, which is why it is crucial in urban planning and policy-making.

The ability of a city to encourage innovation is crucial in attracting investment, retaining talent, and enhancing the life quality for its citizens.

There are various measures to quantify each city's innovation potential and establish comparison rankings between them.

Based on various factors,

research and development capabilities,

cultural assets, available infrastructure, and

the quality of city markets and how connected they are to the world.

In this article, we want to study how the activity carried out by the inhabitants of the city measured through digital traces correlates with innovation indicators.

Introduction

we focus on the activity description based on a decomposition in spatiotemporal city activity patterns and

We study if they allow to predict whether a city is innovative.

Introduction

- Findings
 - It is feasible to predict the innovative character of a city from its activity patterns
 - Identifying indertemination zones improves classification results
 - Synthetic data does not improve the results
 - Random forest provides the best results $AUC=0.71$
 - We identify the features most influential on innovation measure

Dataset

The study analyzes city activity patterns from a social media dataset containing around 32 million geo-tagged urban activities collected from various digital and social platforms over 17 years.

The dataset covers 127 cities worldwide and is available for seven 3-year time slices from 2005 to 2021.

Each city is characterized by a $k \times s$ matrix, where $k = 3$ represents the number of city activity patterns and $s = 7$ represents the number of time slices (weekly patterns)

Innovation Cities Index [1], an annual quantitative index that ranks the most innovative cities globally based on cultural assets, human infrastructure, and networked markets.

The definition of city/town, their location, and respective geographical centers were obtained from the World Cities Database provided by Simplemaps

Dataset

We consider nine features for each city

three features correspond to the average of each activity pattern over time.

a ratio is computed between the average of each activity pattern in the first two time slices against the average of the last two.

a coefficient of variation is computed for each activity pattern that consists of the average of each one divided by its standard deviation.

These features will be used to predict whether or not the city is innovative.

Target variable

The city within the top 50 positions in the City Innovation Ranking will be classified as innovative.

Otherwise, we will classify the city on the way to being innovative or non-innovative.

For the indeterminate dataset, an city between places 50 and 135 in the innovation ranking is labelled as indeterminate.

The city innovation ranking ranges between 1 and 500.

Data augmentation

we compared the performance of a set of state-of-the-art synthetic data generators,

Gaussian Copula, CopulaGAN, CTGAN, and TVAE.

we work with two architectures in the case of CTGAN and TVAE.

Gaussian Copula and CopulaGAN are trained both using the default configuration. Arch A is the default configuration in both cases.

At the same time, Arch B is a setup for the generator with two linear residual layers and the discriminator with two linear layers, both of size (512, 512) for the CTGAN and TVAE Arch B, set hidden layers of (256, 256, 256) for both the encoder and the decoder.

Overall methodology

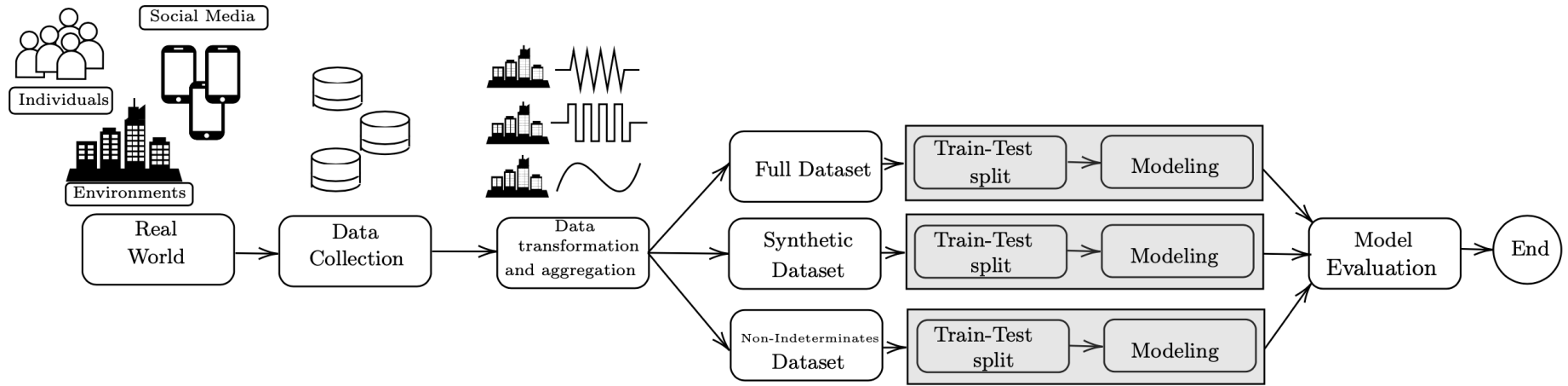


Fig. 1. Proposed methodology for innovative city assessment

Experimental design

In this study, we set $N = 100$, which consists of training 100 models with each data set.

Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Light Gradient Boosting (LGB) are the algorithms used.

For the models trained with the full dataset and the non-indeterminate dataset, an exhaustive search for hyper-parameters is performed using grid search for the LGB model, varying the number of estimators (20, 40), the learning rate (0.01, 0.05, 0.1) and the minimum child samples (2%, 4%).

In the case of the synthetic dataset, 20,000 samples are extracted, and RF and LGB hyper-parameters are optimized using grid search. For both models, the number of estimates is varied (100, 200, 500), and for LGB, the learning rate is varied (0.05, 0.1), and the minimum child samples (2%, 4%), and for RF, the maximum deep is varied (2, 4, 8).

Finally, a non-parametric test (Wilcoxon Test) is used to compare the models.

Results

- Synthetic data similarity to real data

Table 1. Synthetic data generators performance

Synthesizer	Arch	KSTest-mean
Gaussian Copula	A	0.870
Copula GAN	A	0.582
CTGAN	A	0.669
	B	0.712
TVAE	A	0.811
	B	0.856

Results

Table 2. City innovation classification results

Model	Training Data	Model id	AUC	KS	Accuracy	Recall	Precision	F-measure
Logistic Regression	Real Data	LR_r	0.61 ± 0.09	0.35 ± 0.11	0.66 ± 0.07	0.04 ± 0.07	0.12 ± 0.22	0.05 ± 0.09
	Real Data Ind	LR_i	0.63 ± 0.09	0.37 ± 0.12	0.63 ± 0.07	0.18 ± 0.15	0.29 ± 0.19	0.20 ± 0.14
Decision Tree	Real Data	DT_r	0.59 ± 0.07	0.19 ± 0.12	0.64 ± 0.06	0.46 ± 0.15	0.42 ± 0.12	0.42 ± 0.10
	Real Data Ind	DT_i	0.60 ± 0.08	0.22 ± 0.14	0.62 ± 0.07	0.54 ± 0.14	0.42 ± 0.12	0.46 ± 0.10
Light Gradient Boosting	Real Data	LGB_r	0.69 ± 0.08	0.42 ± 0.12	0.70 ± 0.06	0.37 ± 0.21	0.46 ± 0.24	0.39 ± 0.19
	Real Data Ind	LGB_i	0.69 ± 0.07	0.42 ± 0.11	0.68 ± 0.06	0.50 ± 0.21	0.46 ± 0.18	0.45 ± 0.16
	Synthetic Data	LGB_s	0.59 ± 0.08	0.31 ± 0.08	0.71 ± 0.07	0.07 ± 0.07	0.60 ± 0.48	0.13 ± 0.11
Random Forest	Real Data	RF_r	0.69 ± 0.08	0.40 ± 0.11	0.70 ± 0.06	0.35 ± 0.16	0.53 ± 0.18	0.40 ± 0.13
	Real Data Ind	RF_i	0.71 ± 0.08	0.42 ± 0.11	0.70 ± 0.06	0.51 ± 0.16	0.52 ± 0.15	0.50 ± 0.11
	Synthetic Data	RF_s	0.60 ± 0.08	0.32 ± 0.09	0.70 ± 0.07	0.07 ± 0.07	0.49 ± 0.42	0.12 ± 0.11

Results

$$\frac{AUC_{row} - AUC_{column}}{AUC_{column}}$$

Table 3. Model comparison based on AUC results

AUC	LR_r	LR_i	DT_r	DT_i	LGB_r	LGB_i	LGB_s	RF_r	RF_i	RF_s
LR_r	*	-2.5%	*	*	-10.6%	-11.4%	4.7%	-11.4%	-14.2%	*
LR_i	2.5%	*	6.7%	*	-8.3%	-9.1%	7.4%	-9.2%	-12.1%	4.6%
DT_r	*	-6.3%	*	*	-14.1%	-14.9%	*	-14.9%	-17.6%	*
DT_i	*	*	*	*	-11.9%	-12.7%	*	-12.7%	-15.5%	*
LGB_r	11.9%	9.1%	16.4%	13.5%	*	*	17.1%	*	-4.1%	14.2%
LGB_i	12.9%	10.1%	17.5%	14.5%	*	*	18.2%	*	-3.2%	15.2%
LGB_s	-4.5%	-6.9%	*	*	-14.6%	-15.4%	*	-15.4%	-18.1%	-2.5%
RF_r	12.9%	10.1%	17.5%	14.6%	*	*	18.2%	*	-3.2%	15.2%
RF_i	16.6%	13.7%	21.4%	18.3%	4.2%	3.3%	22.1%	3.3%	*	19.0%
RF_s	*	-4.4%	*	*	-12.4%	-13.2%	2.6%	-13.2%	-16.0%	*

Results

Table 4. Feature Importance

Feature Name	Importance
Activity Pattern_1_cv	25.6%
Activity Pattern_2_R	14.2%
Activity Pattern_1_R	13.6%
Activity Pattern_0_mean	10.1%
Activity Pattern_2_mean	9.6%
Activity Pattern_0_R	7.8%
Activity Pattern_2_cv	7.3%
Activity Pattern_1_mean	6.3%
Activity Pattern_0_cv	5.6%

Results

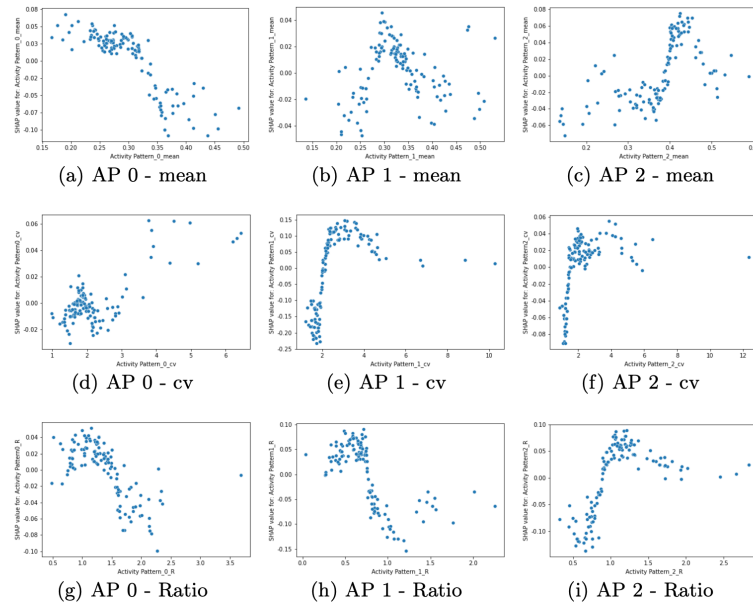


Fig. 2. Feature Importance Analysis using SHAP Values. Each subfigure shows the predictor along with their corresponding SHAP Values.

Conclusions

In this study, we associate characteristics of a city obtained from the activity of its inhabitants with structural characteristics such as a city's ability to innovate.

To meet this objective, we use the activity pattern decomposition of each city and its respective innovation ranking.

Our proposal models this challenge as a classification problem.

For the training, three approaches are used: working with the complete data set, increasing the information with synthetic data, and removing indeterminates.

The results show that training without indeterminates achieves the best results when training a random forest model, achieving better results on all the proposed metrics.

This model performs well in determining if a city will be innovative, with an AUC of 0.71 and a KS of 0.42.

Additionally, we present the features that most influence this probability