Universidad del País Vasco
Euskal Herriko Unibertsitatea

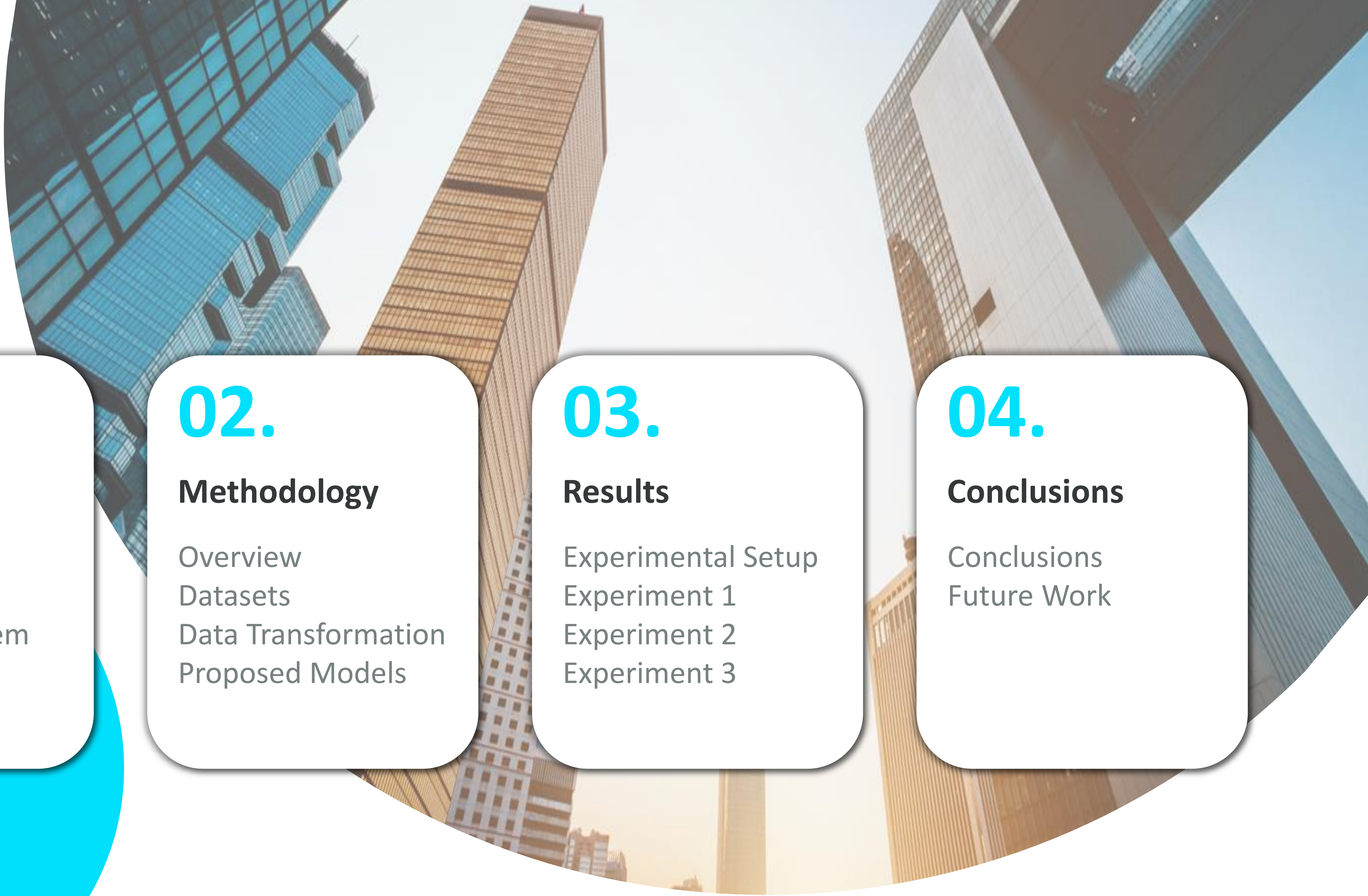
On the use of multi-sensor digital traces to discover spatio-temporal human behavioral patterns

By Ricardo Luis Muñoz Cancino

Department of Computer Science and Artificial Intelligence
PhD Thesis Defense to obtain the degree of Doctor in Computer Science

PhD Advisors:
Dr. Manuel Graña Romay
Dr. Sebastián A. Ríos Pérez

Contents



Motivation

Motivation

Motivation


中国电信
5:28
7月14日 星期二
庚子年五月廿四


BANCA BT TRANSILVANIA
Antreprenor
credit
4140 1234 5678 9010
4119
EXP 12/25
VISA
Business
DAN POPESCU
SC DANPO SRL
Business
VISA


08/12 MON
10:09
8190


oyster
Transport for London
Issued subject to conditions – see over

Motivation

Geotagged Urban Activities

The things people do, where they do, and when they do them

Motivation

The most important uses of human digital traces

- Urban planning and infrastructure management
- In the event of a natural disaster. Aid efforts and plan the response
- Quantify the effect of climate change
- The spread and contagion of biological viruses
- The response and impact of a terrorist attack
- Management, characterizing, and forecast of vehicle traffic
- Public transportation management and public policies
- Forecasting people's crowd flows

Gaps

In this thesis work, we will address three main gaps identified.

Algorithms

Patterns validation

Time-dependent patterns

Research Problem

General Objective

The main objective of this thesis is to extend the general knowledge of the use of digital traces to study the patterns of human activity and its interaction with the environment.

Specific Aims

Aim 1: Formalizing the human behavioral patterns validation

Research Questions: Is it possible to reduce the dependence on expert evaluators to identify human behavioral patterns? What should be the main metrics to evaluate and choose between a set of patterns?

Broader Implications: Formalizing the evaluation and selection of human behavioral patterns will facilitate research on this topic because it is no longer dependent on experts in the geographical areas of study.

Specific Aims

Aim 2: Modeling human behavioral patterns

Research Questions: Is it possible to detect human behavioral patterns? What is the best way to represent digital traces to detect these patterns? What relationship do these patterns have with the geographic area of data collection?

Broader Implications: Surveys are the most widespread way to learn about people's mobility patterns. A method that allows us to delve into human behavioral patterns through passively collected information will reduce costs and increase the speed and frequency with which insights can be obtained.

Specific Aims

Aim 3: Temporal and spatial human behavioral patterns

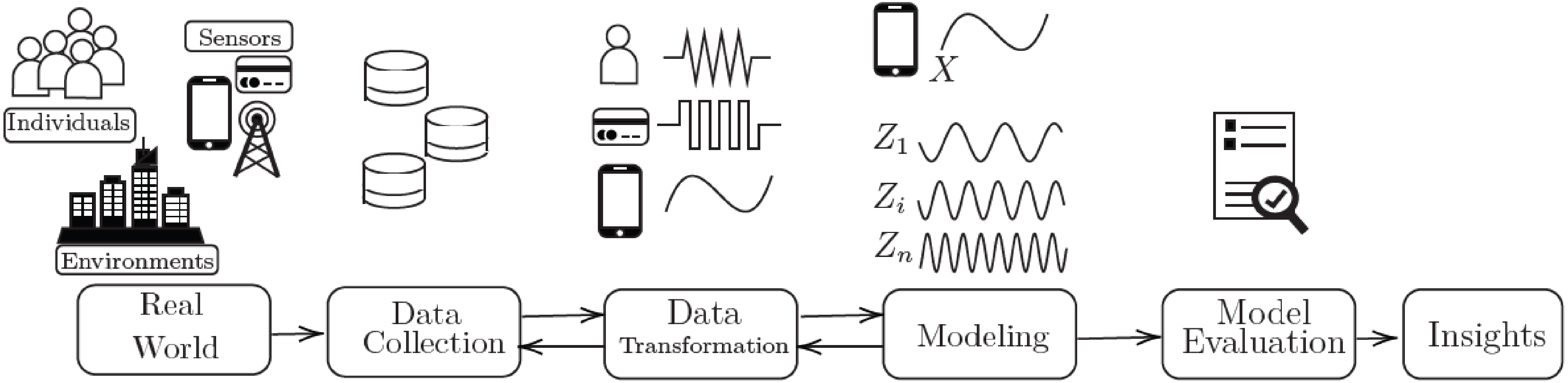
Research Questions: Are the patterns detected stable over time? How can we incorporate the temporal dimension in identifying human behavioral patterns?

Broader Implications: Human behavior changes, as does their interaction with their neighborhoods and cities. For this reason, having a methodology that allows us to observe these changes in behavior is of great help for decision-making in public policies and all other uses given to this information.

Contents

Methodology

A general approach



Data Collection

Data sources used in this research

Telecom

Call detail records

880 million records
3 million individuals
77 days

Banking

Credit card purchases

190 million records
80,000 terminals
3 years

Social Media

Geotagged urban activities

32 million records
140 cities
17 years

Data Collection

Privacy protection and ethical guidelines

Data sources do not reveal any individual's identity or personal information.

Telecom and banking datasets were used at the antenna and point of sales levels, respectively, containing only aggregate information

Online activity data was obtained from public datasets aggregated at the city level, without considering the user or individual.

The final data produced by this research does not compromise customers' privacy and cannot leak any personal private information.

The Telecom Dataset

Call Details Records

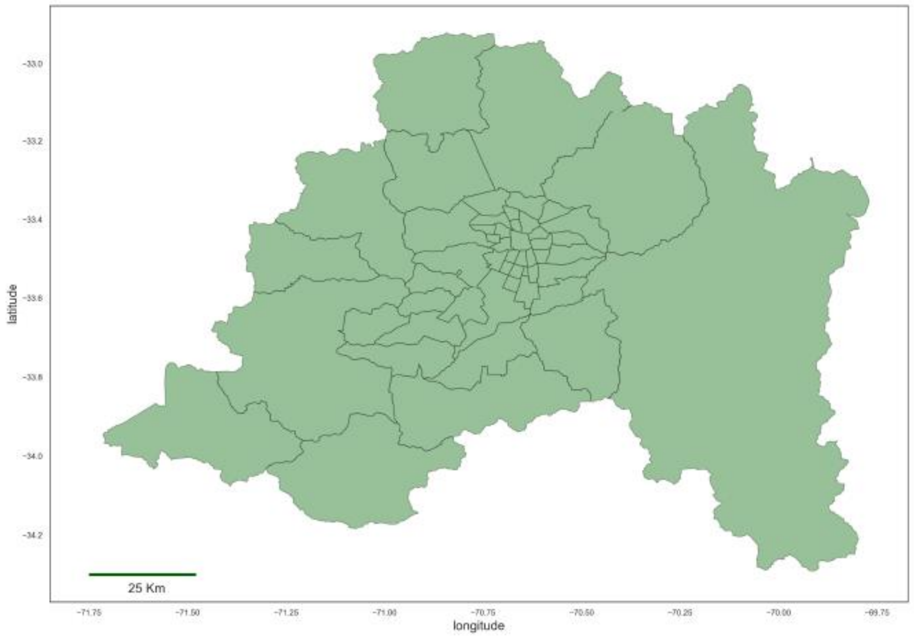
Base transceiver station

The Telecom Dataset

Call Details Records

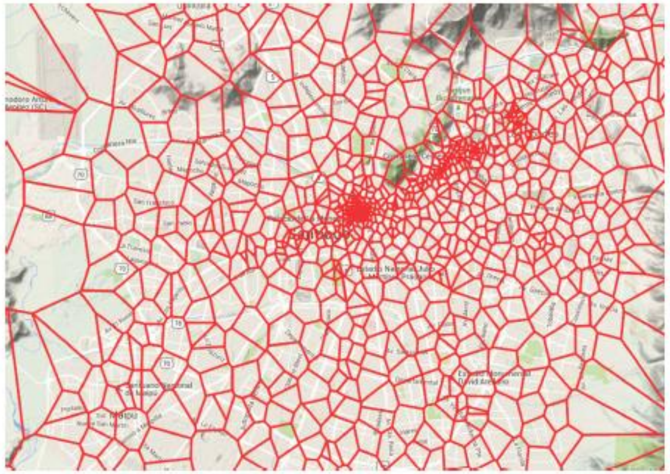
Base transceiver station

Caller Phone Number	Callee Phone Number	Caller BTS	Callee BTS	Timestamp	Call duration (seconds)
0012542872	0042478890	BTS00001	BTS00234	2022-04-10 23:15:18	230
0012542872	0056978425	BTS00041	BTS00934	2022-08-17 08:25:13	37
...	...	...	...	...	...
...	...	...	...	...	...
0085967423	0012457784	BTS02477	BTS00065	2022-09-30 09:55:23	108



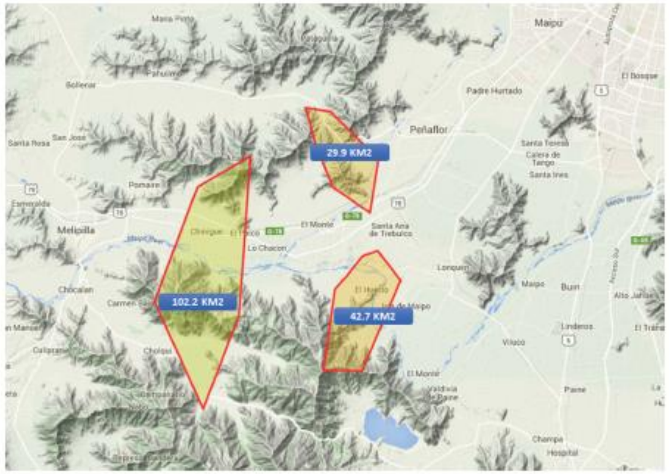
The Telecom Dataset

Study Area

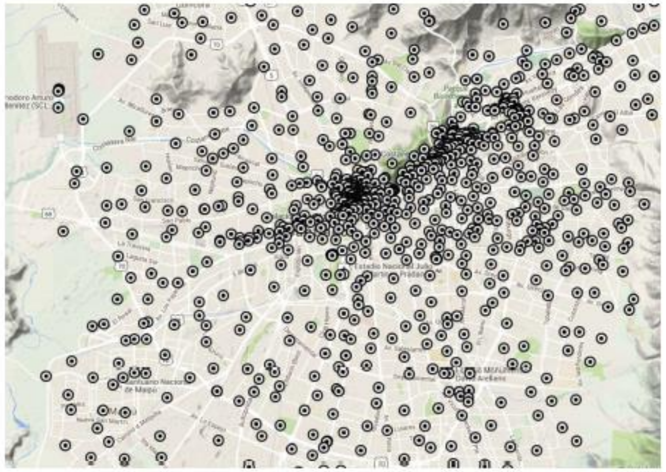


Santiago Tessellation

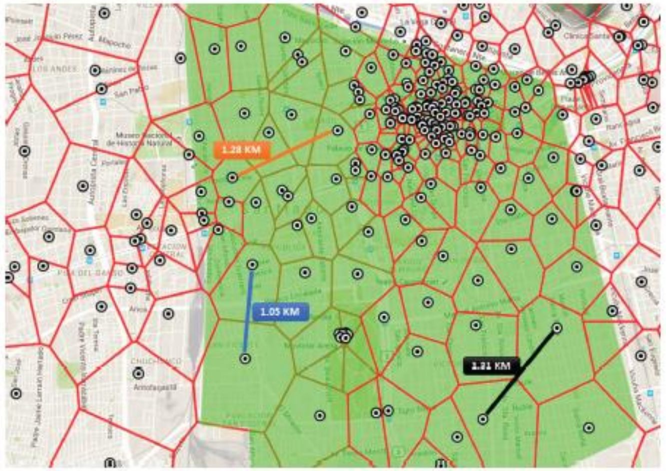
Urban Areas



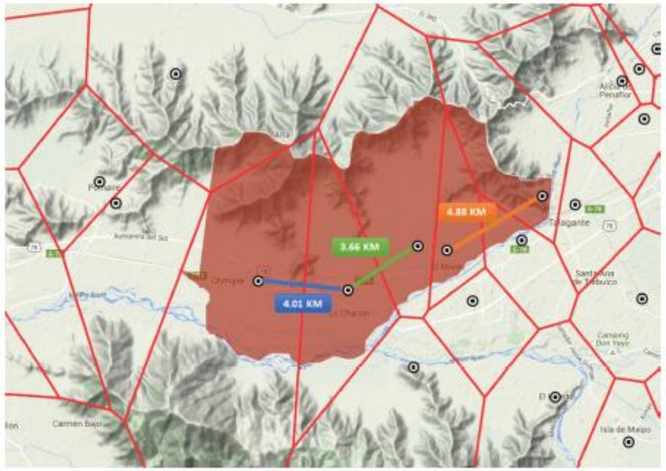
Rural Areas



Santiago BTS Positions



Urban Areas

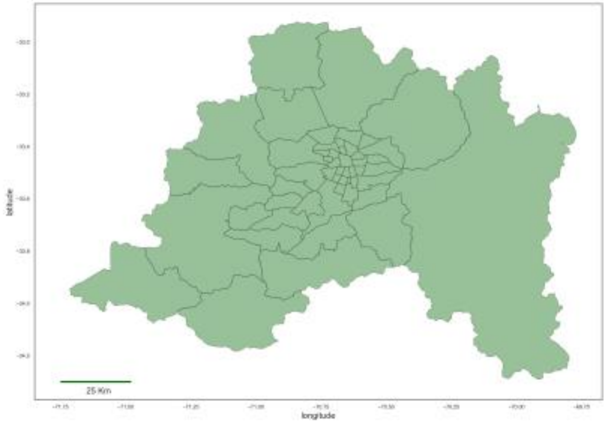


Rural Areas

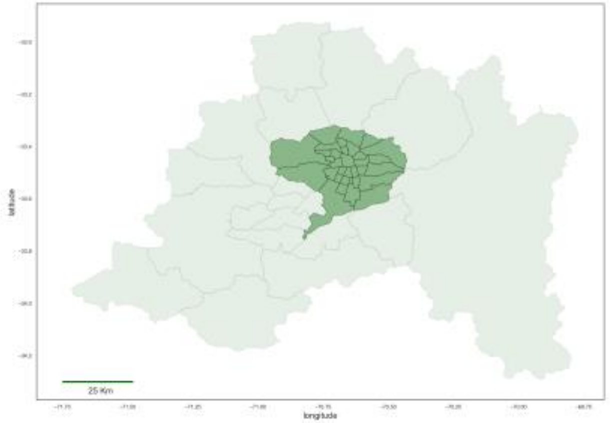
The Banking Dataset

Credit and Debit card records

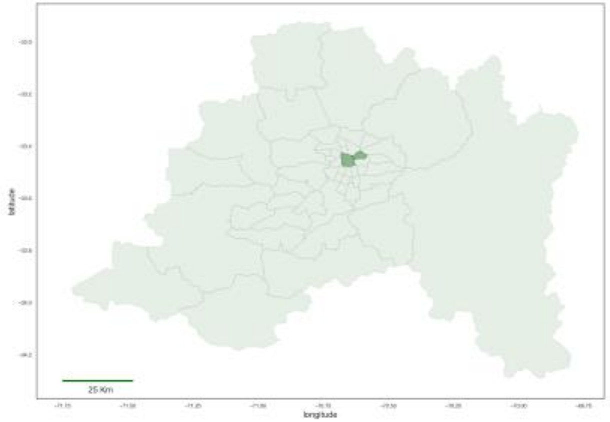
Point of sale terminal



Greater Santiago



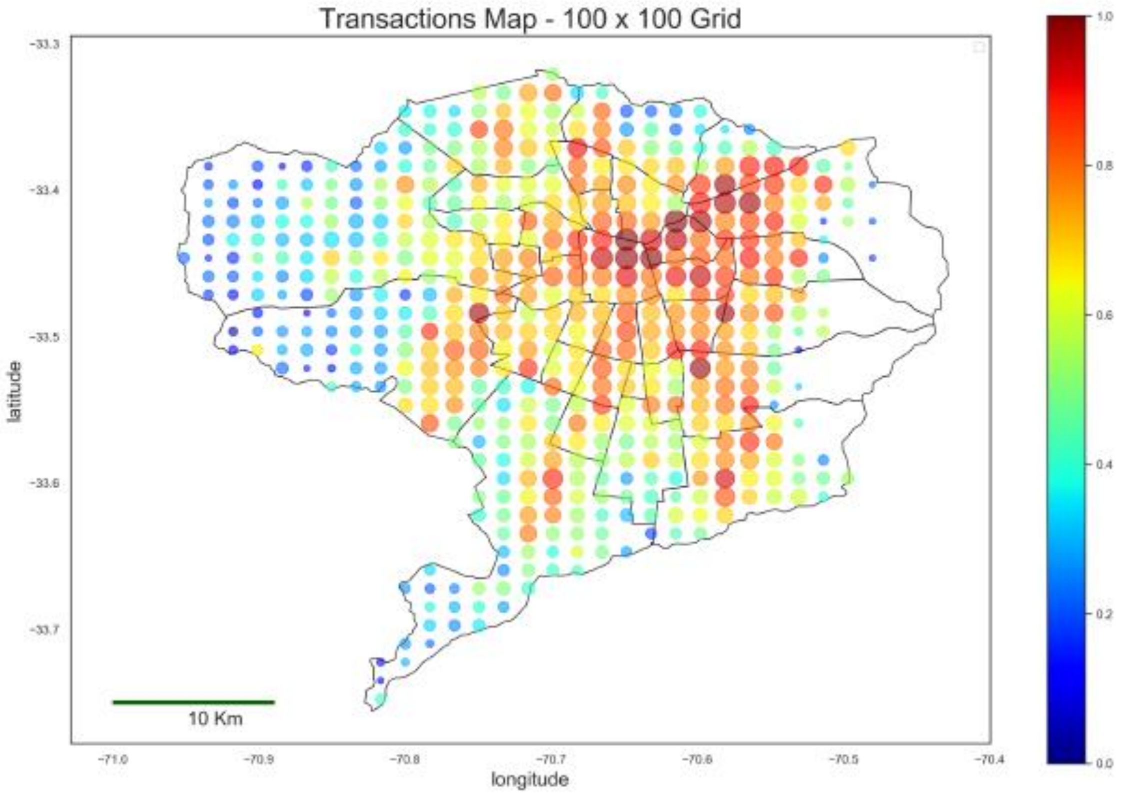
Santiago City

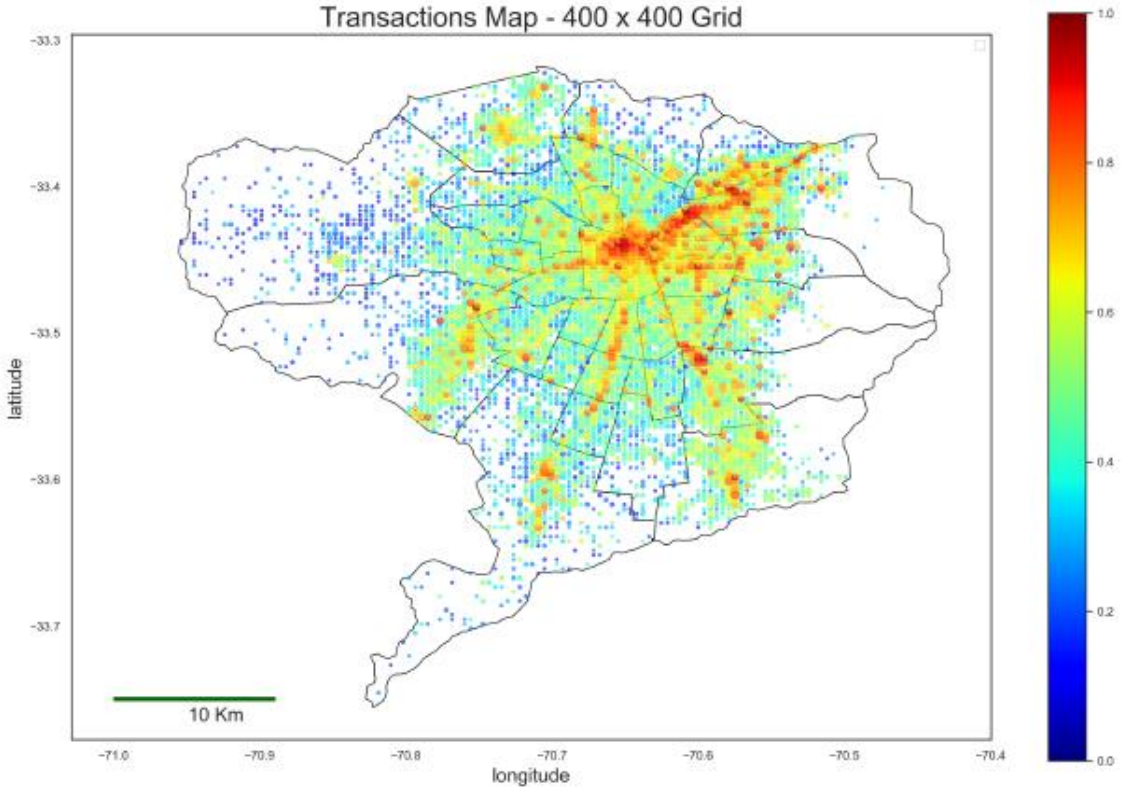


Downtown Santiago

The Banking Dataset

Credit and Debit card records





The social media dataset

Geo-tagged urban activity records

Source	Events	#Cities	#Year	Min Year	Max Year
Brightkite checkins	1,639,399	107	3	2008	2010
Foursquare checkins	7,515,201	107	6	2010	2015
	109,9826	3	2	2012	2013
GeoTagged Images	4,998,865	130	8	2005	2012
GeoTagged Tweets	2,041,262	10	4	2007	2010
	187,802	130	2	2020	2021
	47,337	7	1	2020	2020
	184,547	130	1	2020	2020
	2,604,233	136	1	2020	2020
Gowalla checkins	1,992,082	107	2	2009	2010
Weeplaces checkins	4,176,673	107	7	2005	2011
Yelp checkins	5,695,209	2	12	2010	2021

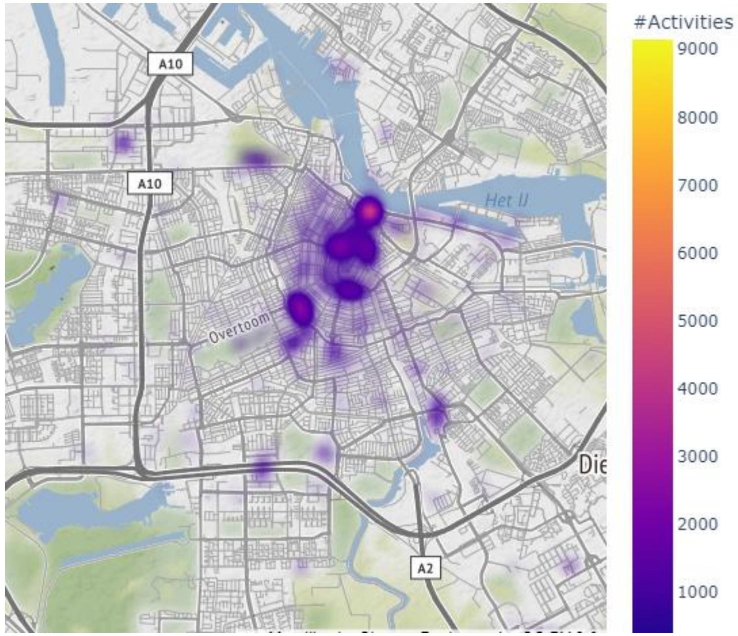
The social media dataset

Study Area

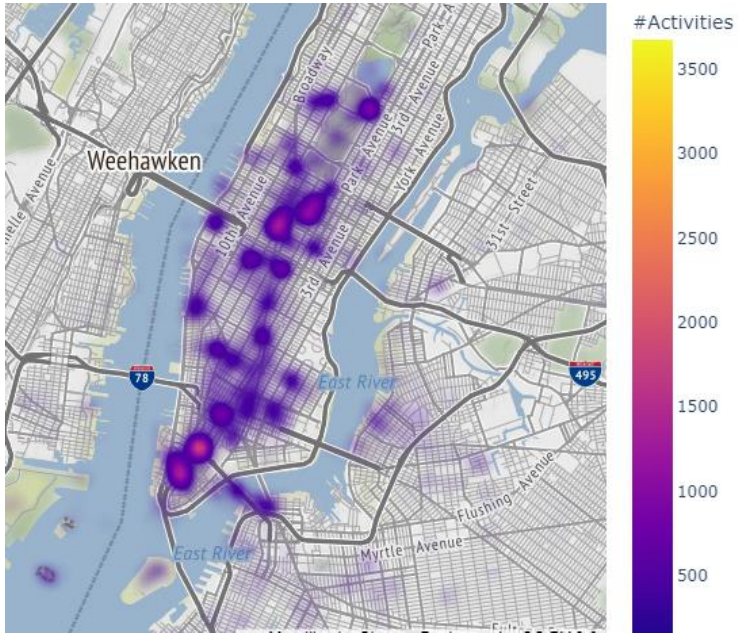
Cities considered in the analysis – more than 1MM inhabitants or capitals

The social media dataset

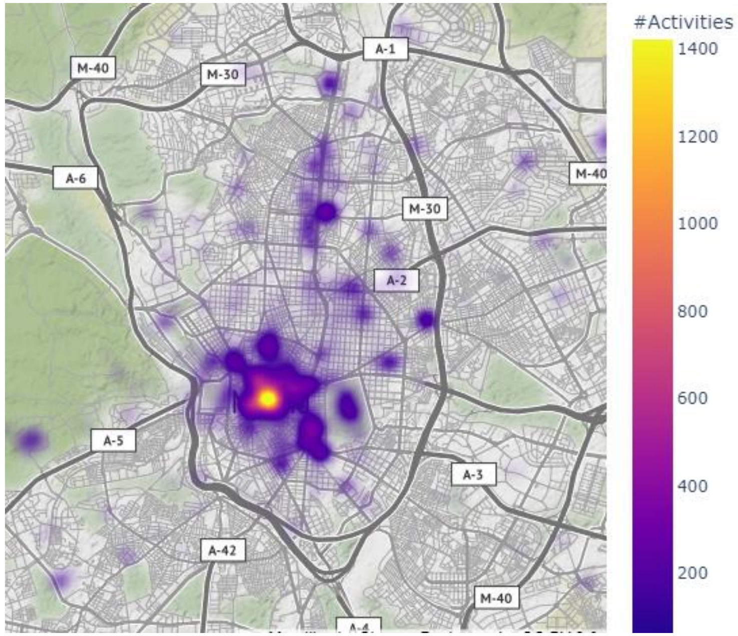
Study Area



Amsterdam



Manhattan



Madrid

Contents

Dataset representation

Activity Patterns

$\mathbf{XP}^s$ Raw activity pattern

XP_t^s Activity Block: the number of geo-tagged events or digital traces (sensor s, period t)

$$\mathbf{XP}^s = (XP_1^s, XP_2^s, \ldots, XP_T^s)(t \in T)$$

T is the set of activity blocks

$$\mathbf{card}(T) = 168$$

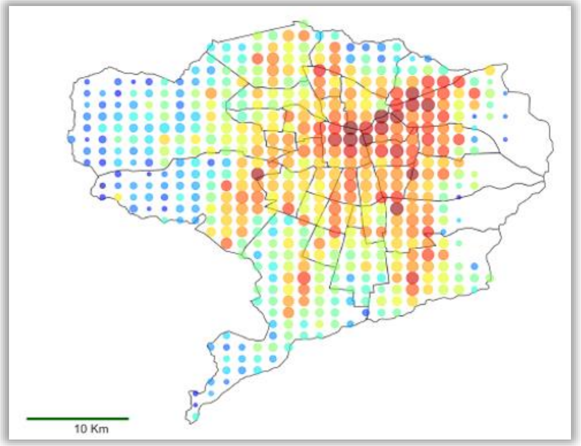
$\mathbf{AP}^s$ Normalized activity pattern

$$AP_t^s = \frac{XP_t^s}{\sum_{t \in T} XP_t^s}$$

Data transformation

And spatial aggregation

Voronoi Cell



Grid Cell

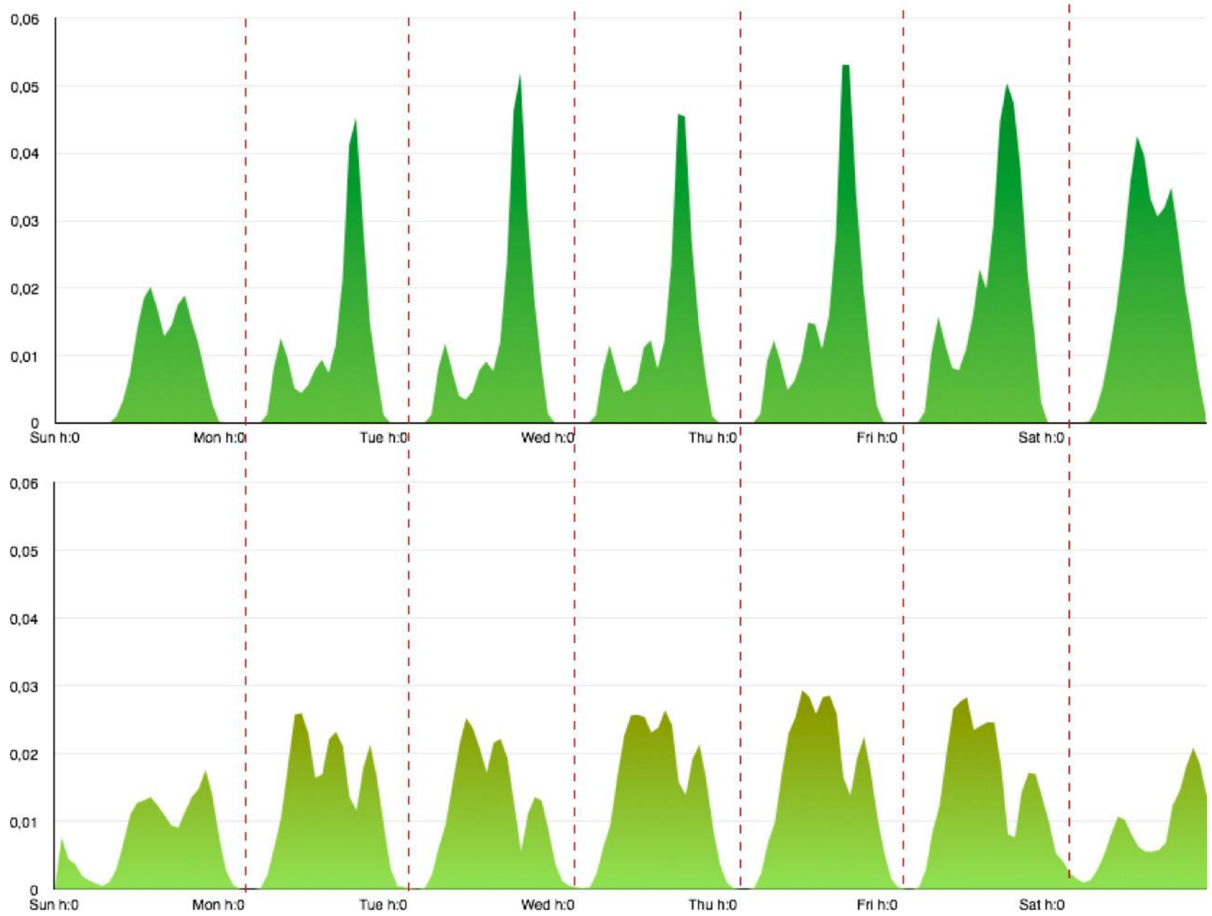
City

$$\mathbf{XP}_a = \frac{1}{\mathbf{card}(H_a)} \sum_{h \in H_a} \mathbf{XP}^h \qquad \mathbf{AP}_a = \frac{1}{\mathbf{card}(H_a)} \sum_{h \in H_a} \mathbf{AP}^h$$

H_a is the set of sensor inside the aggregated area a

Data transformation

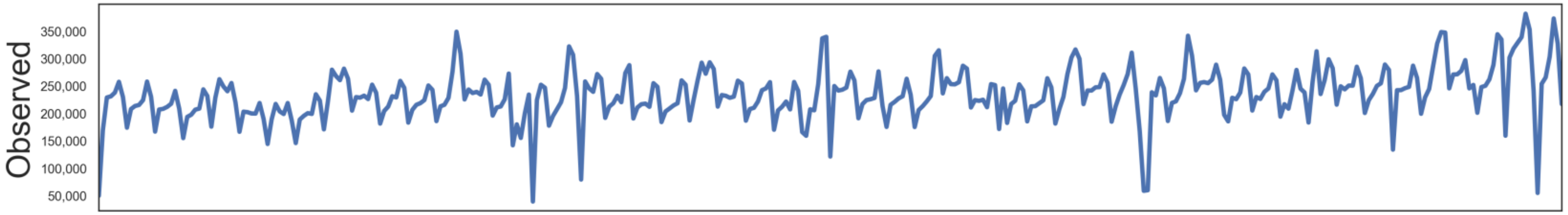
And spatial aggregation



Activity Patterns example using the telecom dataset

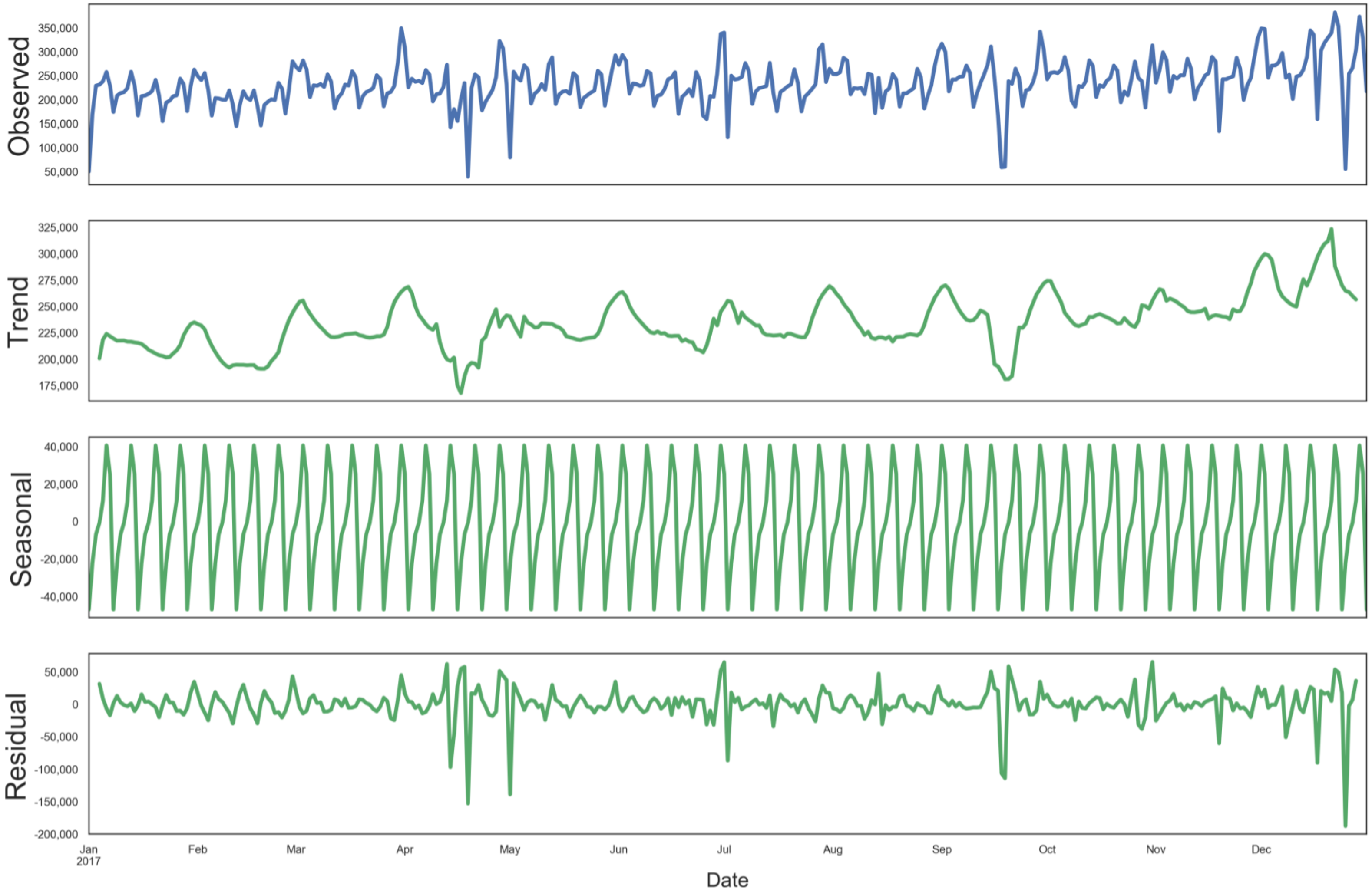
Activity pattern duration

Determination of activity pattern duration

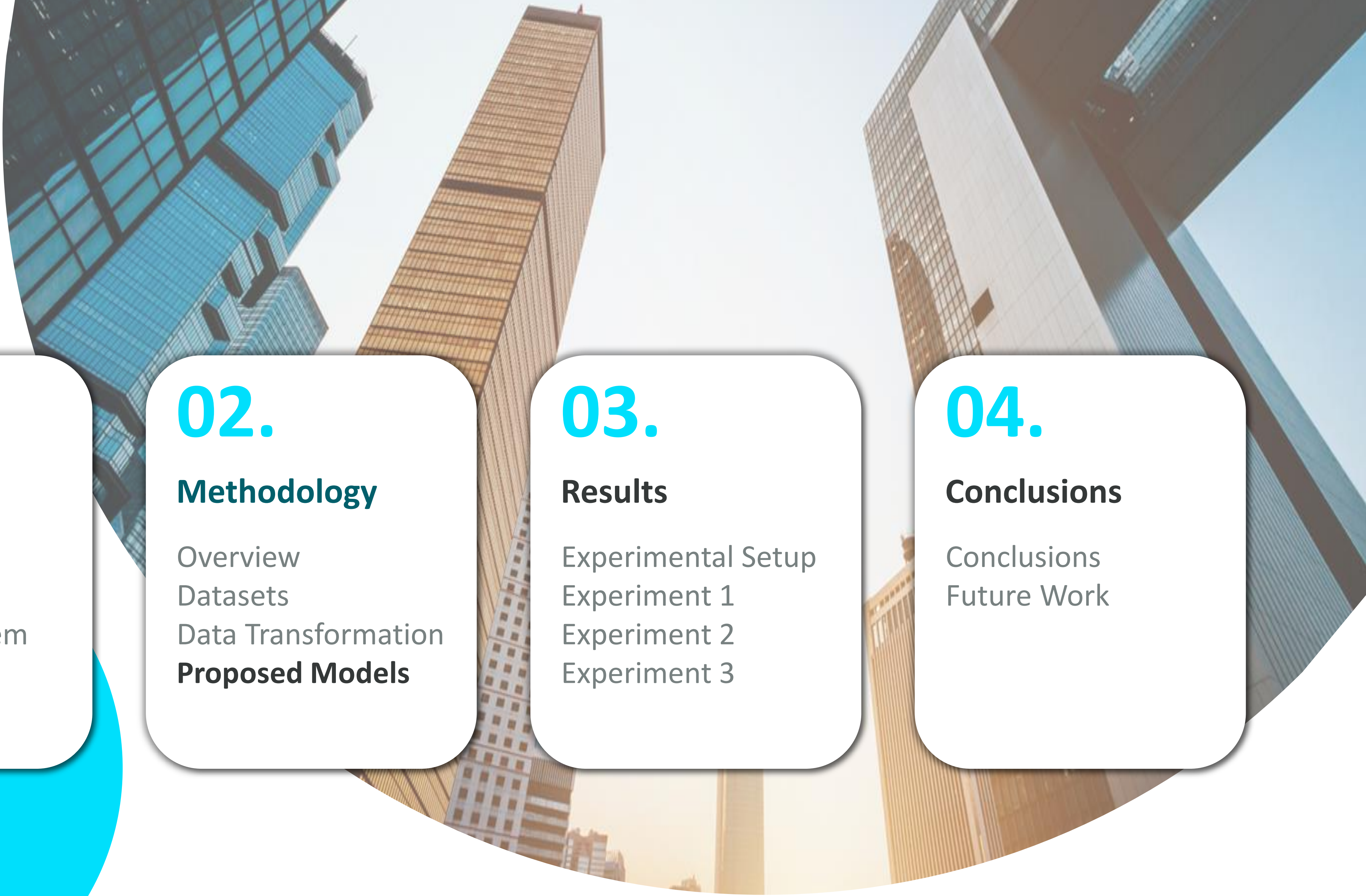


Activity pattern duration

Determination of activity pattern duration



Contents

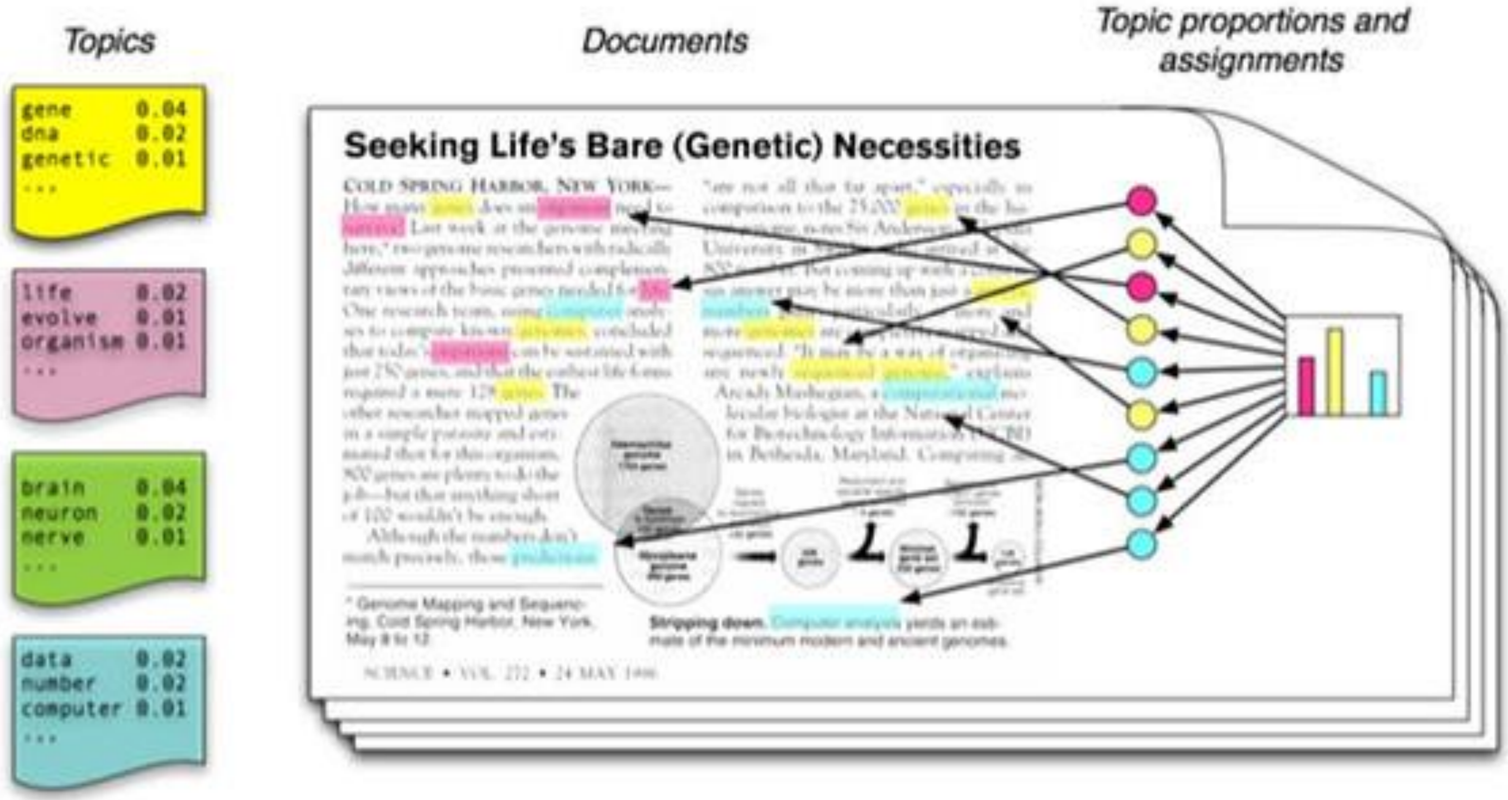


Static Topic Modeling

Latent Dirichlet Allocation

The Idea:

- Each document is a mixture of corpus-wide topics
- Each topic is a distribution over words
- Each word is drawn from one of the topics



Source: Blei, D. M. (2012), Probabilistic topic models.

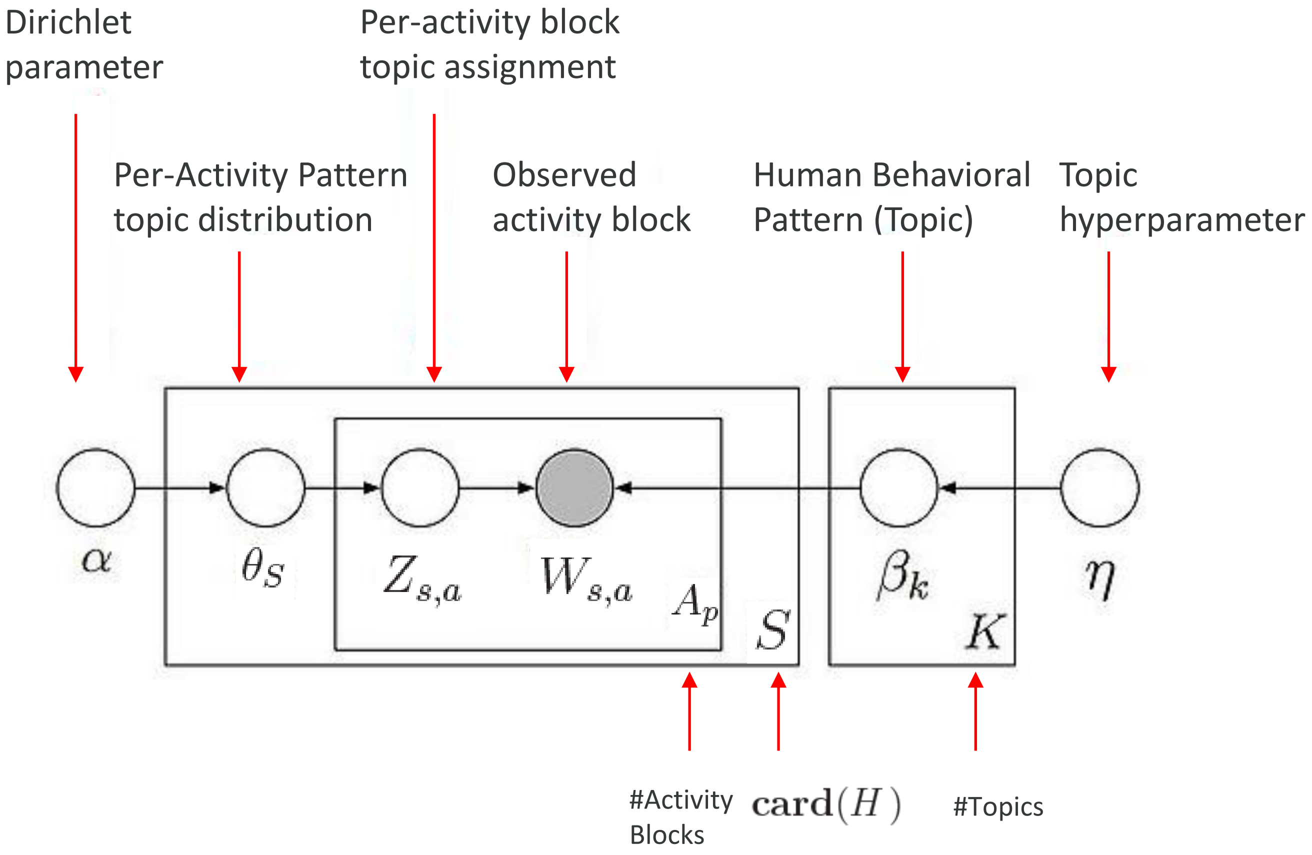
Static Topic Modeling

Latent Dirichlet Allocation using geo-tagged digital traces

The main objective of this model is to learn Human Behavioral Patterns from digital traces data distribution by inferring latent topics

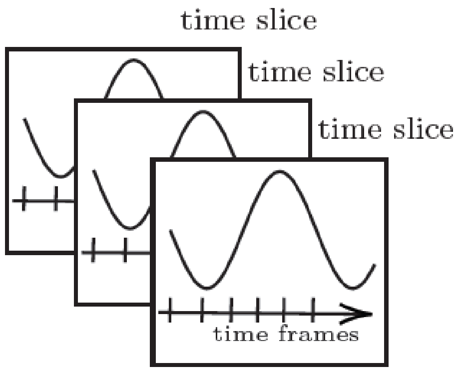
$$p(\theta, z, \mathbf{a}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta)p(a^s|z_s, \beta)$$

The exact computation of the posterior in a Bayesian approach is intractable. Instead, Gibb Sampling is used.



Dynamic Topic Modeling

Dynamic topic models using geo-tagged digital traces

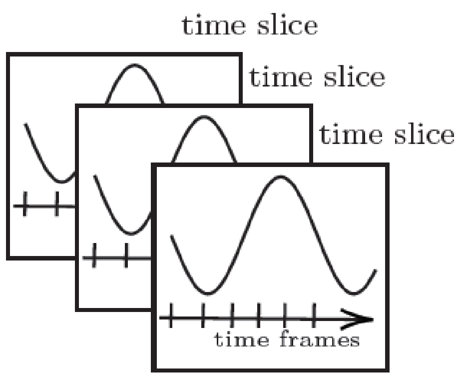


The Idea:

- Divide the dataset by time slice
- Start with a separate LDA for each year
- Then add a dependence of each time slice on the previous one

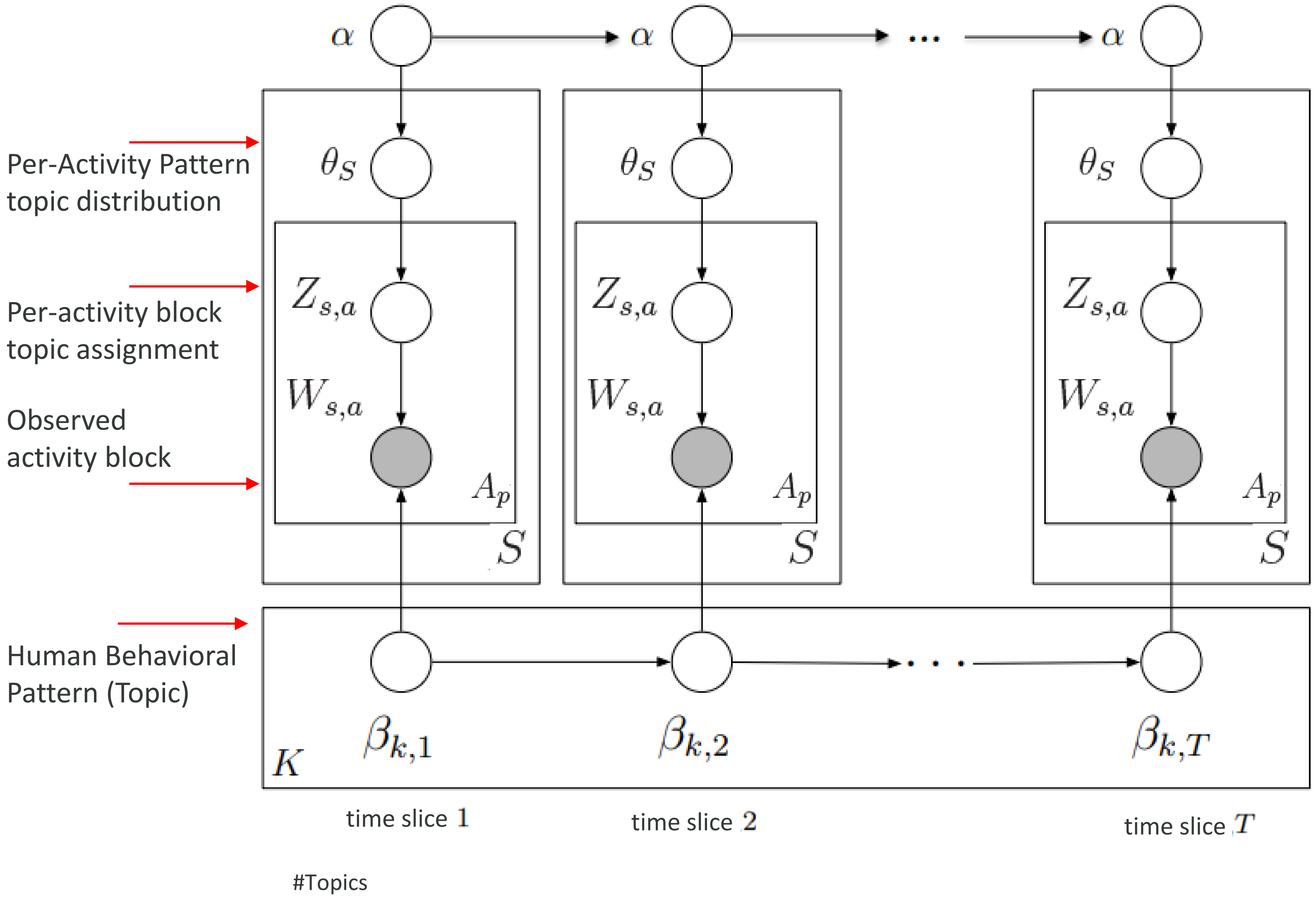
Dynamic Topic Modeling

Dynamic topic models using geo-tagged digital traces



The Idea:

- Divide the dataset by time slice
- Start with a separate LDA for each year
- Then add a dependence of each time slice on the previous one

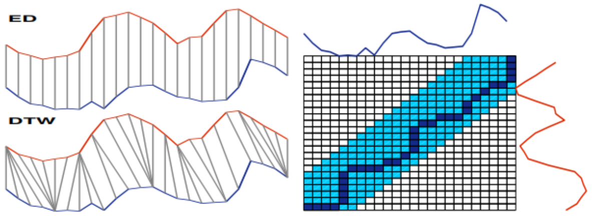


Traditional Algorithms

To detect human activity patterns

K-means

K-Shape



Source: Paparrizos et al. (2016), k-Shape: Efficient and Accurate Clustering of Time Series

Human Behavioral Patterns (number)

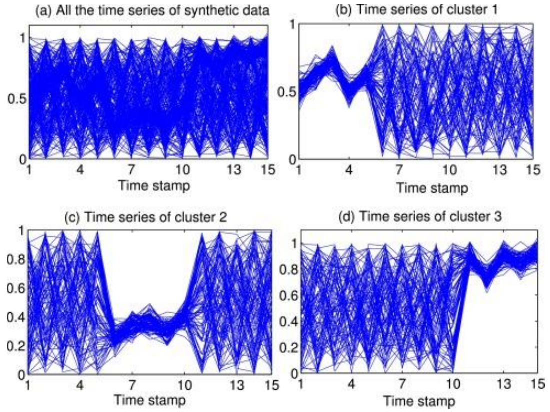
Sensor's cell

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

Resulting cells partition (Voronoi, grid, cities)

Centroid for pattern i

Time series K-Means



Source: Huang et al. (2016), Time series k-means: A new k-means type smooth subspace clustering for time series data

Human behavioral patterns validation

Guidelines on the quality of the topics discovered

Proposed Metrics:

- Intratemporal Similarity
- Intertemporal Stability
- Topic Consistency
- Topic Smoothness

Intratemporal Similarity

Human behavioral patterns validation

The Idea:

One of the main expected results is that the topics or patterns describe different activities carried out in a city. In this way, we expect that human behavioral topics be as dissimilar as possible between them.

Number of Patterns

Different patterns

$$IntraSim(\mathcal{K}) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{2}{K(K-1)}\sum_{k=0}^{K-1}\sum_{l=0}^{K-1}\mathbb{1}_{k\neq l}\cdot sim(AT_{s,k}, AT_{s,l})$$

Time-slice partition

Similarity function

Human Behavioral Pattern

Intertemporal Stability

Human behavioral patterns validation

The Idea:

When the human behavior is analyzed over time, it is expected that the behavior measured from urban activities will not change overnight. We will measure these gradual changes in individuals' behavior from the changes between the same human behavioral pattern in adjacent time-slices.

Number of time-slices

Consecutive time-slices

$$InterSta(\mathcal{K}) = \frac{1}{K} \sum_{k=0}^{K-1} \frac{1}{2|\mathcal{S}| - 2} \sum_{s=0}^{|\mathcal{S}|-1} \sum_{u=0}^{|\mathcal{S}|-1} \mathbb{1}_{|s-u|=1} \cdot sim(AT_{s,k}, AT_{u,k})$$

Number of Patterns

Time-slice partition

Similarity function

Human Behavioral Pattern

Topic Consistency

Human behavioral patterns validation

The Idea:

The human behavioral is based on their daily routines, which is why there is a certain regularity in the activities carried out on working days and during the weekend. This consistency can be observed in the empirical results obtained in multiple studies.

Weight

Weight

Different patterns

$$TC(AT_{s,k}) = \frac{5}{7} \sum_{i,j \in weekdays} \mathbb{1}_{i \neq j} \cdot sim(AT_{s,k,i}, AT_{s,k,j}) + \frac{2}{7} \sum_{i,j \in weekend} \mathbb{1}_{i \neq j} \cdot sim(AT_{s,k,i}, AT_{s,k,j})$$

Time-slice partition and pattern

Similarity function

Human Behavioral Pattern

Topic Smoothness

Human behavioral patterns validation

The Idea:
Human behavior while interacting with the urban infrastructure is carried out continuously and without significant restrictions that suddenly limit all activity. For this reason, patterns with smooth changes are preferred to those with high volatility during the day.

Time-slice partition and pattern

Number of activity blocks

$$TS(AT_{s,k}) = \sqrt{\frac{1}{T-2}\sum_{i=1}^{T-1}(d_{s,k,i} - \bar{d}_{s,k})^2}$$

$$d_{s,k,i} = AT_{s,k,i+1} - AT_{s,k,i}$$

Consecutive activity blocks

$$\bar{d}_{s,k} = \frac{1}{T-1}\sum_{i=1}^{T-1} d_{s,k,i}$$

Mean difference vector

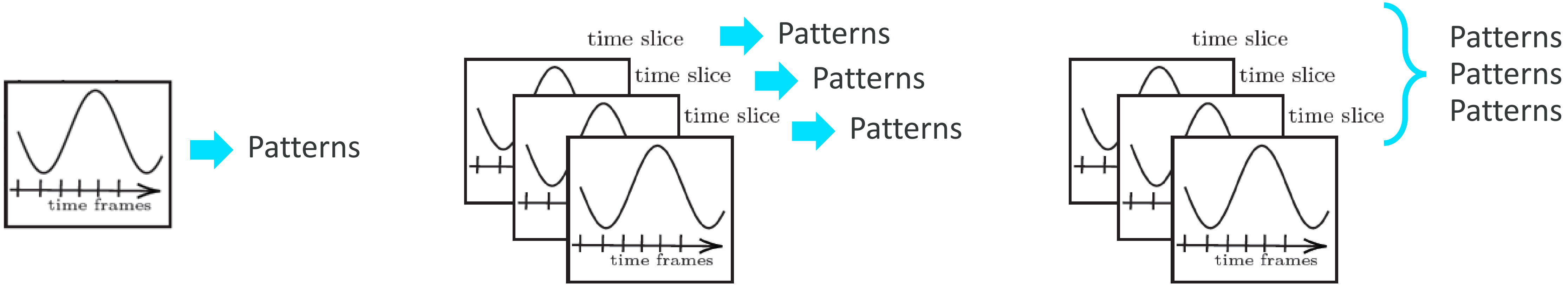
Contents

Experimental Setup

The proposed experimental setup to answer the research questions

In order to answer the research questions, three studies are proposed, each based on the previous results, to provide a complete and rigorous assessment of the proposed methodology.

- **Experiment 1:** Spatial human behavioral patterns
- **Experiment 2:** Spatiotemporal human behavioral patterns: Multiple static models
- **Experiment 3:** Spatiotemporal human behavioral patterns: Model-embedded patterns



Experiment 1 **Experiment 2** **Experiment 3**

Experiment 1: Spatial human behavioral patterns

Single model using the whole dataset

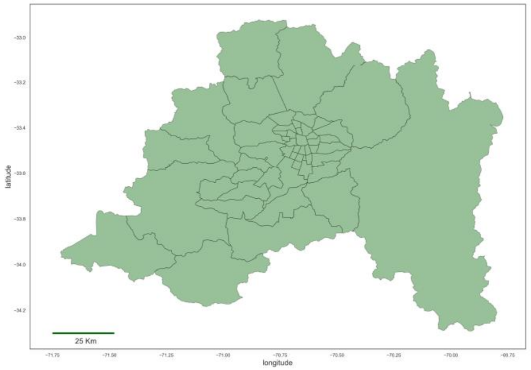
This experiment was designed to address Aim 2 of this thesis, determine if it is possible to detect human behavioral patterns from digital traces and propose alternatives to traditional algorithms for this task.

Telecom

Call detail records

880 million records
3 million individuals
77 days

Dataset

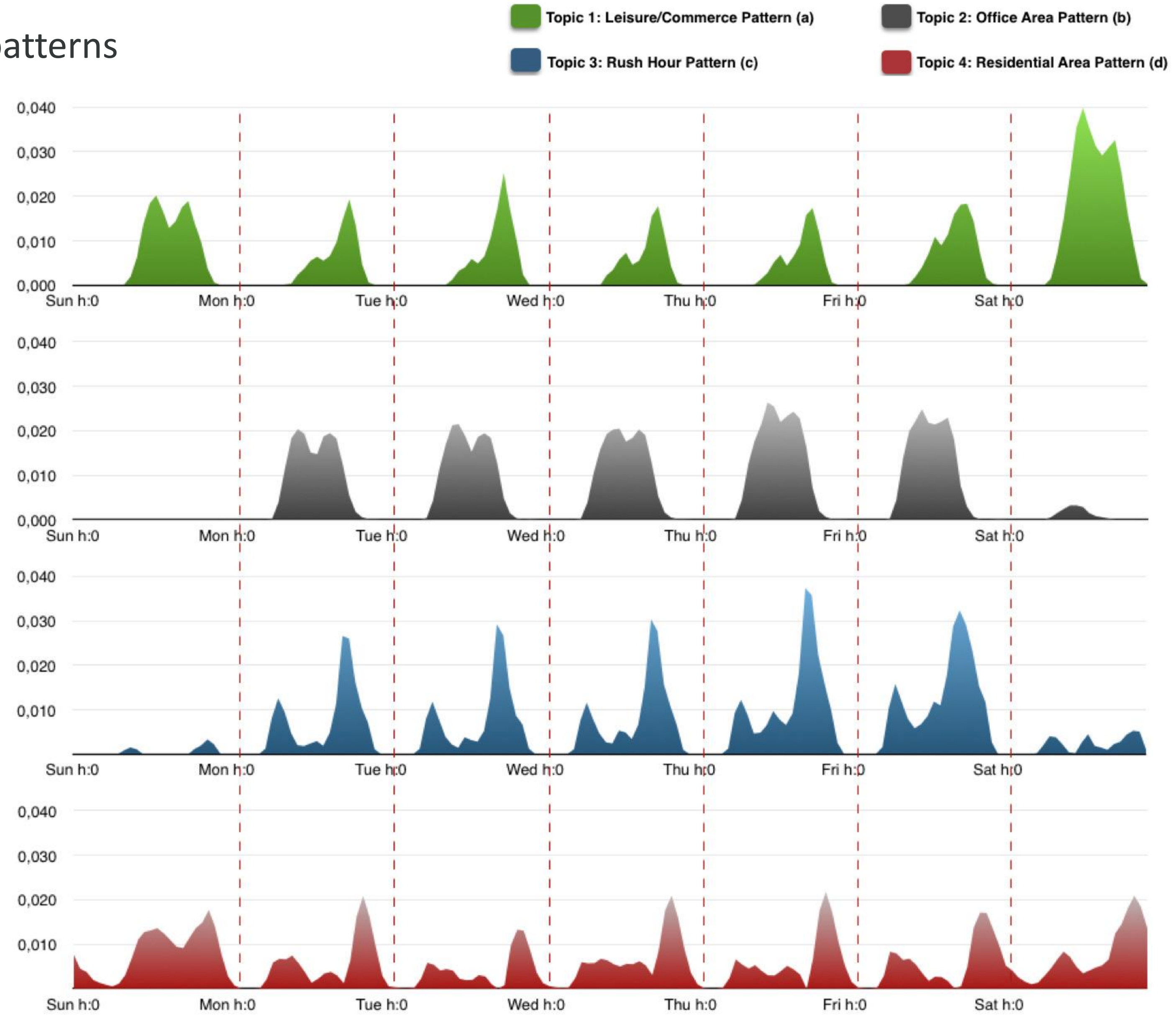


Study Area

Spatial Aggregation

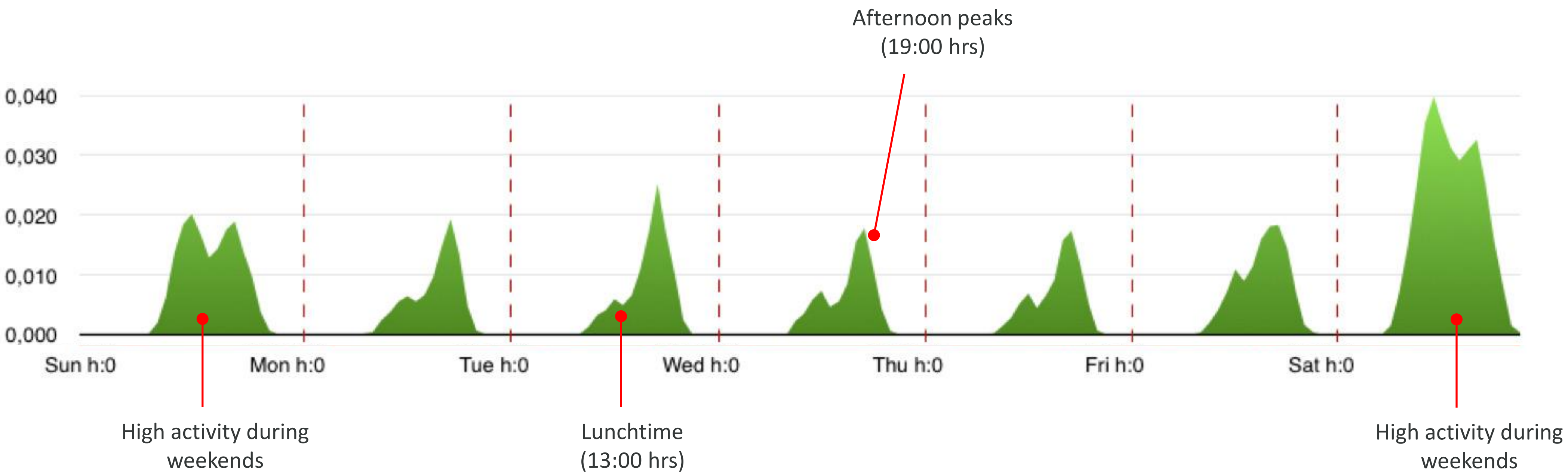
Experiment 1: Results

Spatial human behavioral patterns



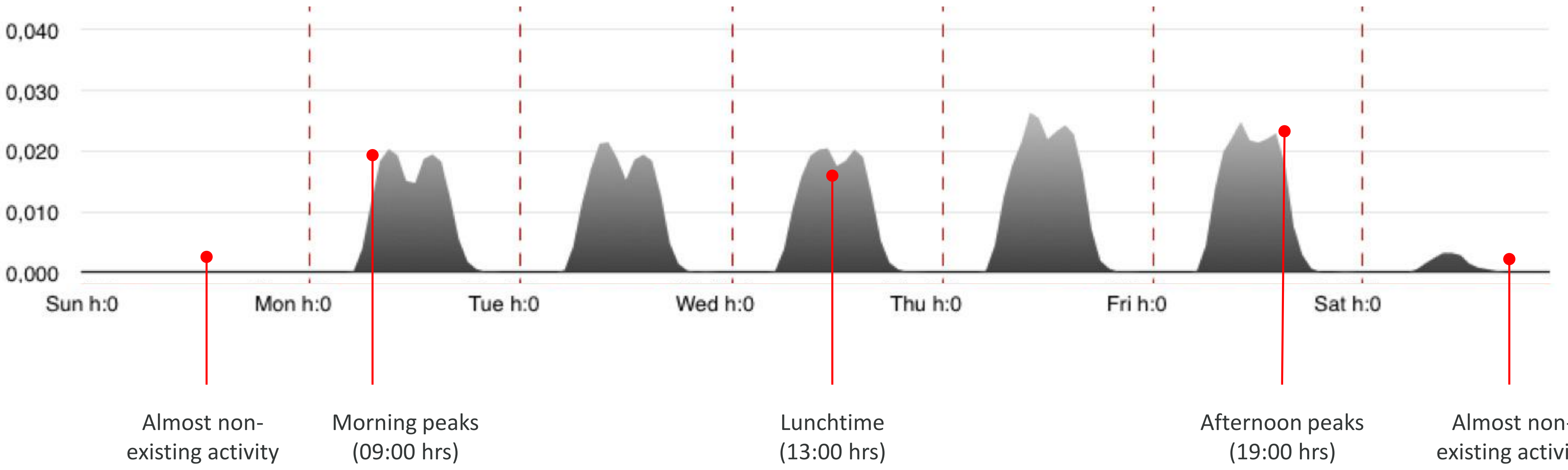
Experiment 1: Results

Topic 1: Leisure/Commerce Pattern



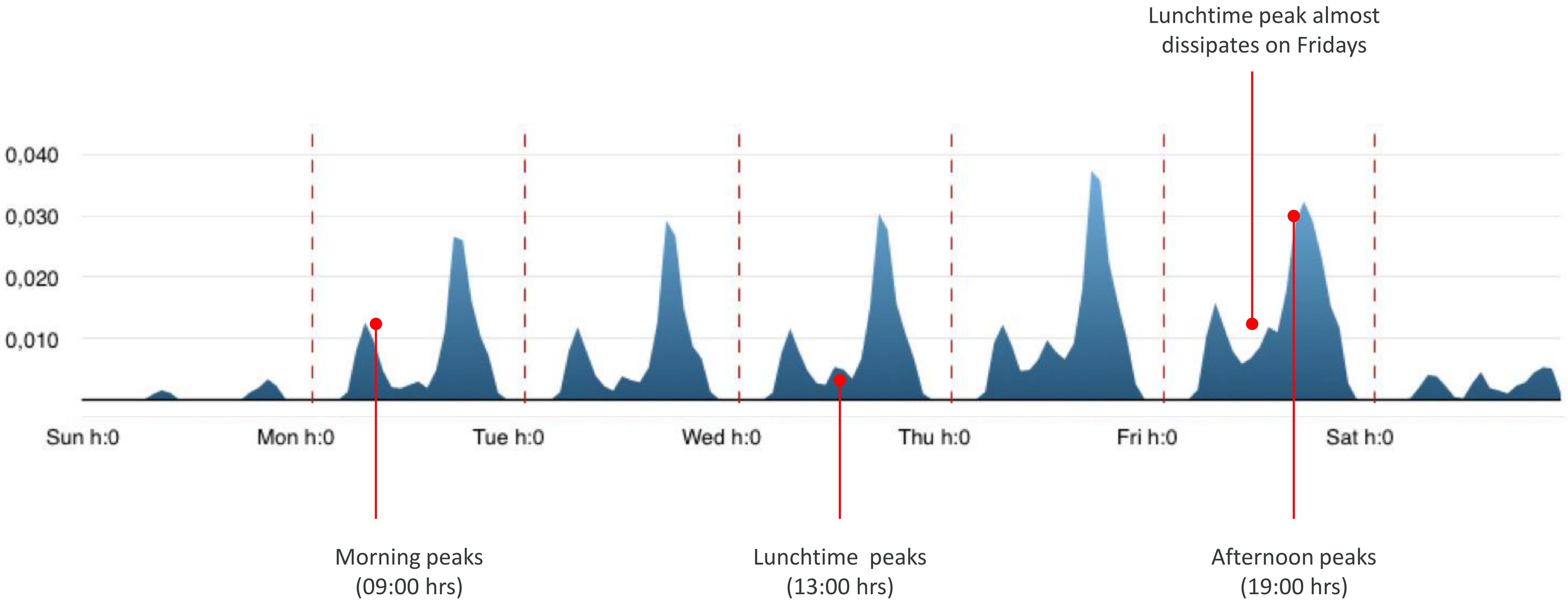
Experiment 1: Results

Topic 2: Office Area Pattern



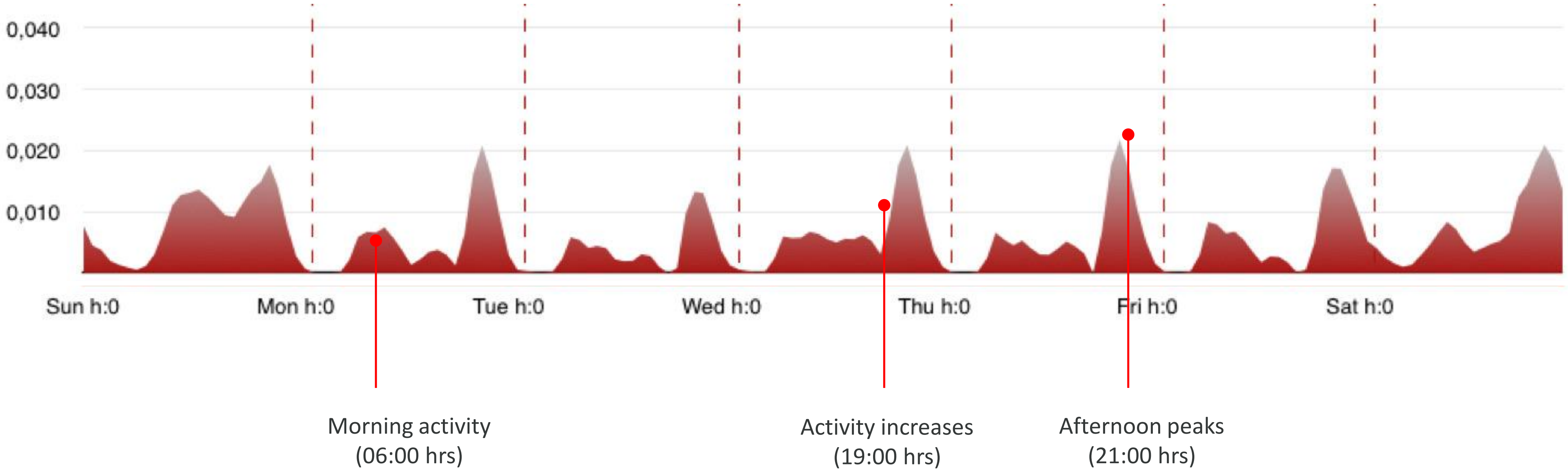
Experiment 1: Results

Topic 3: Rush Hour Pattern



Experiment 1: Results

Topic 4: Residential Area Pattern



Experiment 1: Stability

Spatial human activity patterns stability

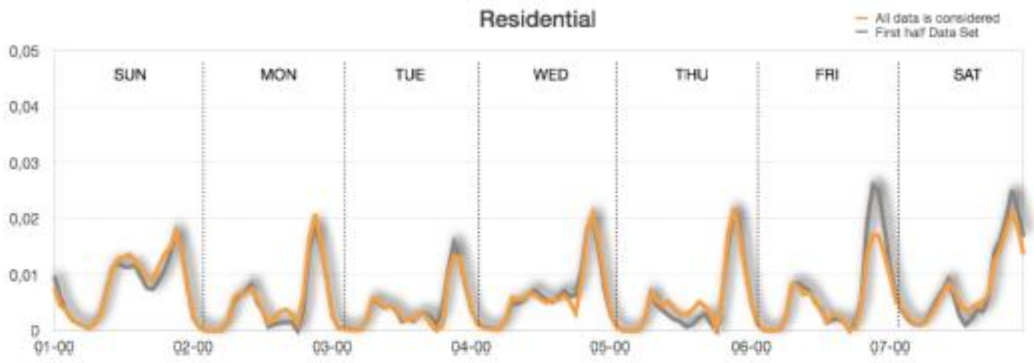
In order to analyze the stability of the patterns discovered using the methodology presented in this work, we divided the telecom dataset into two equal data subsets. The first dataset contains calls made between 2013-04-18 and 2013-05-26, and the second one between 2013-05-27 and 2013-04-07.

To quantify the pattern stability under different datasets, we used Cosine Similarity. This measure is defined as follows

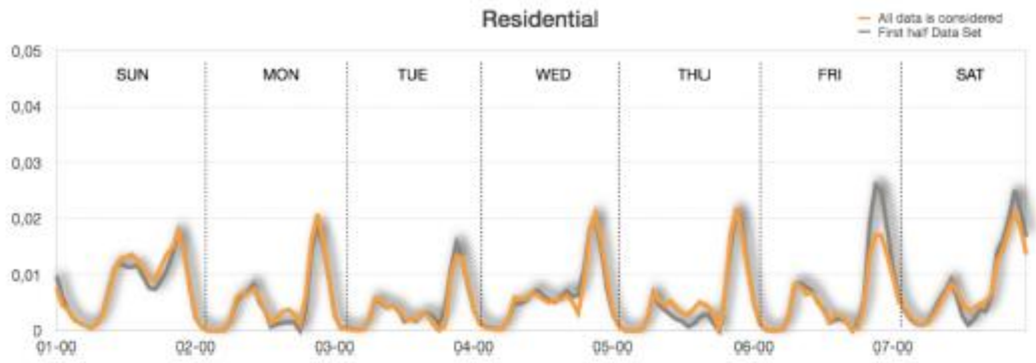
$$COS(AP_b, AP_c) = \frac{\sum_{i=1..N} A_i^b \cdot A_i^c}{\sqrt{\sum_{i=1..N}(A_i^b)^2} \cdot \sqrt{\sum_{i=1..N}(A_i^c)^2}}$$

Experiment 1: Stability

Spatial human activity patterns stability

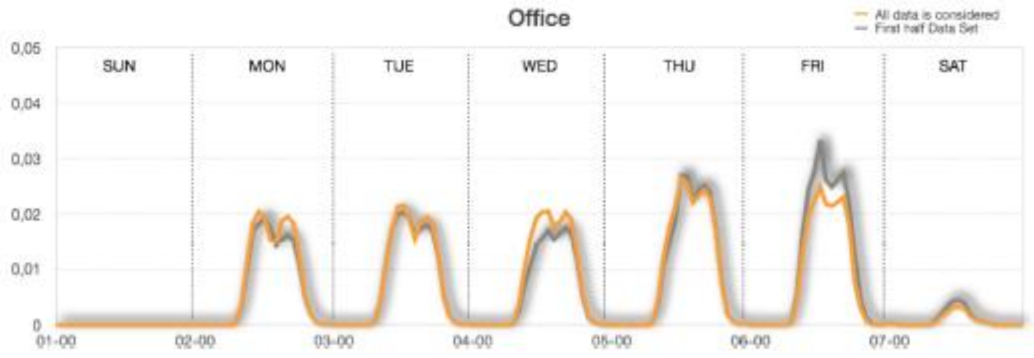


(a) First Half Dataset

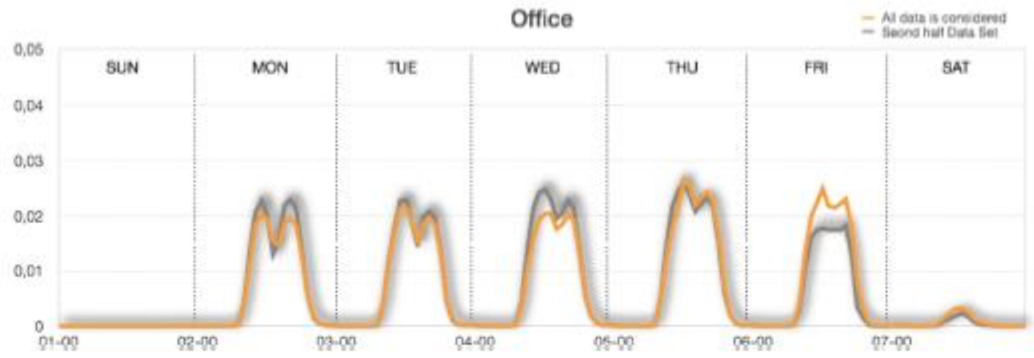


(b) Second Half Dataset

Residential Pattern Stability



(a) First Half Dataset

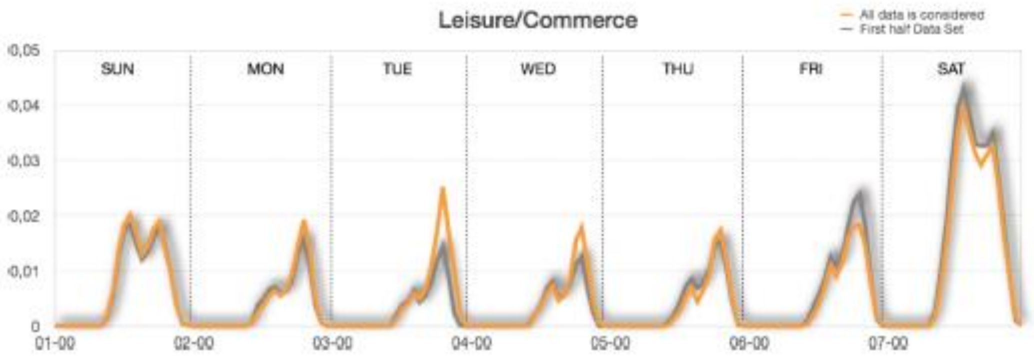


(b) Second Half Dataset

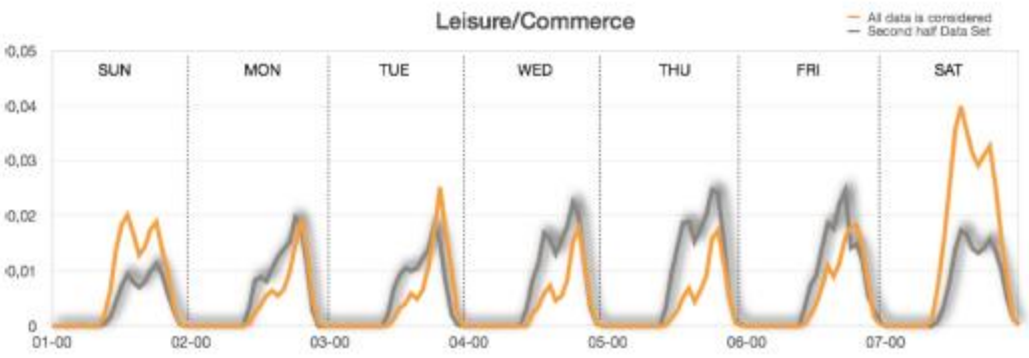
Office Areas Pattern Stability

Experiment 1: Stability

Spatial human activity patterns stability

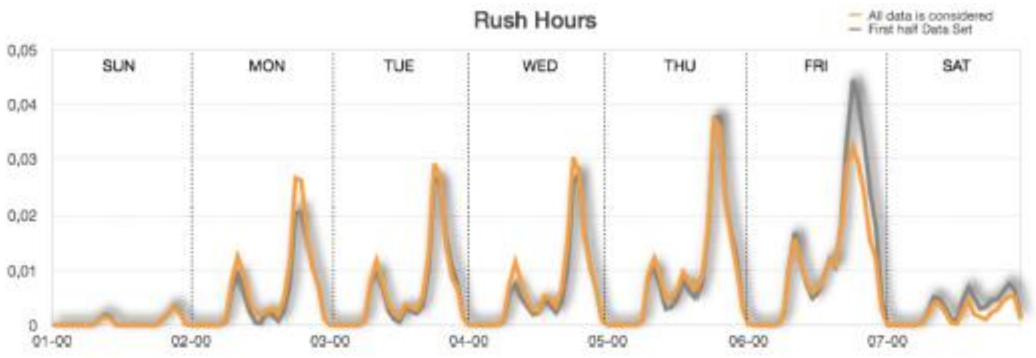


(a) First Half Dataset

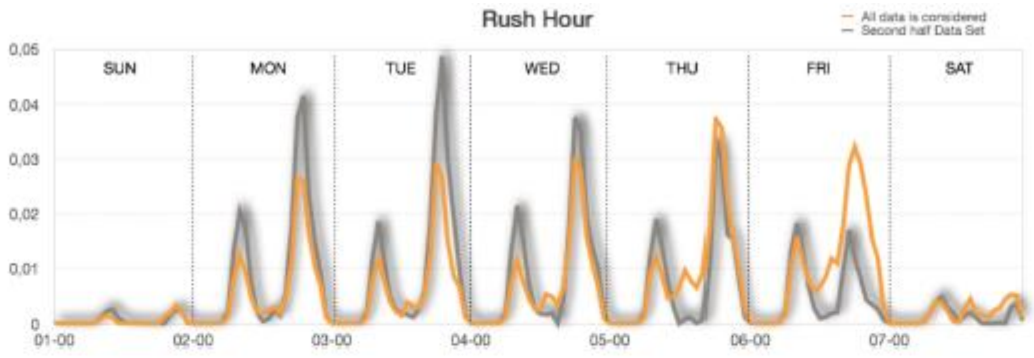


(b) Second Half Dataset

Leisure/Commerce Pattern Stability



(a) First Half Dataset



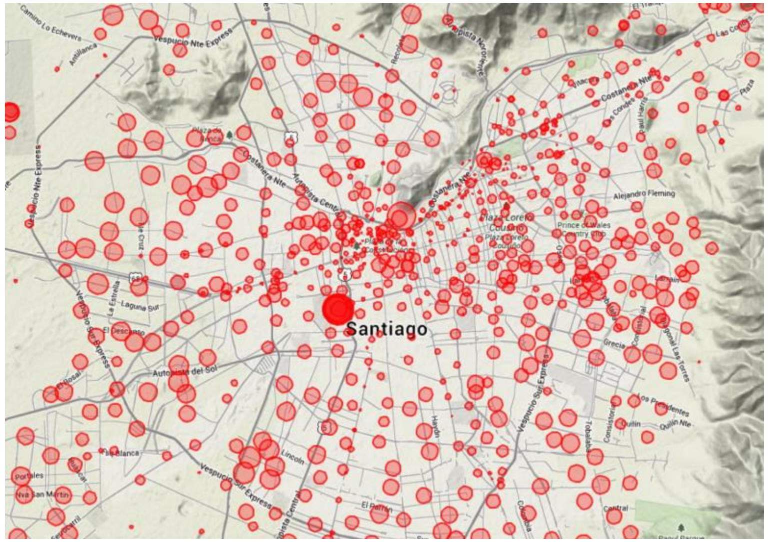
(b) Second Half Dataset

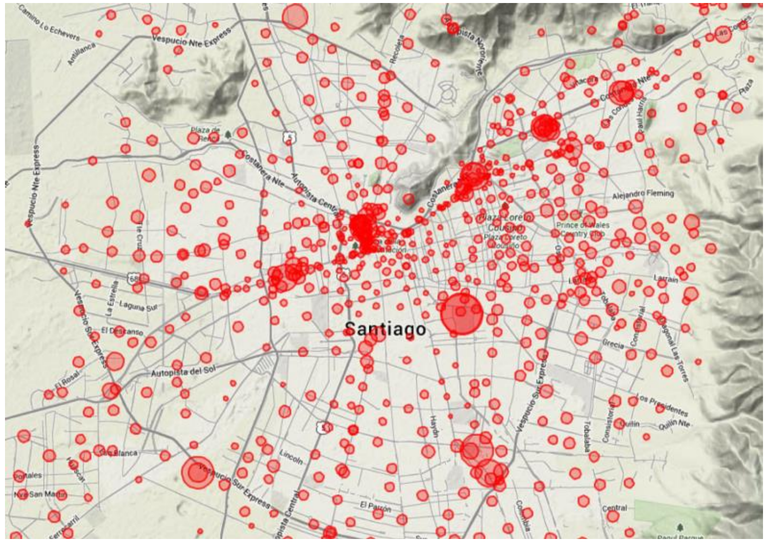
Rush Hour Pattern Stability

Experiment 1: A spatial comparison

Geographical representation of human behavioral patterns

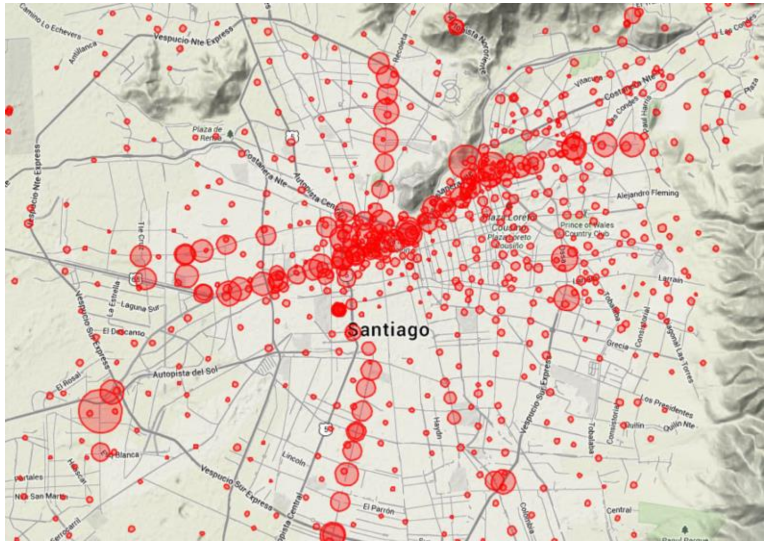
Residential

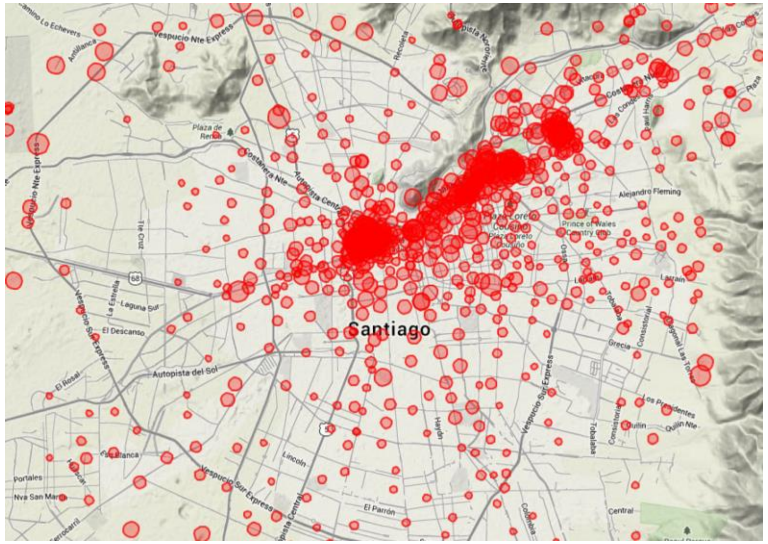




Leisure-Commerce

Rush Hour



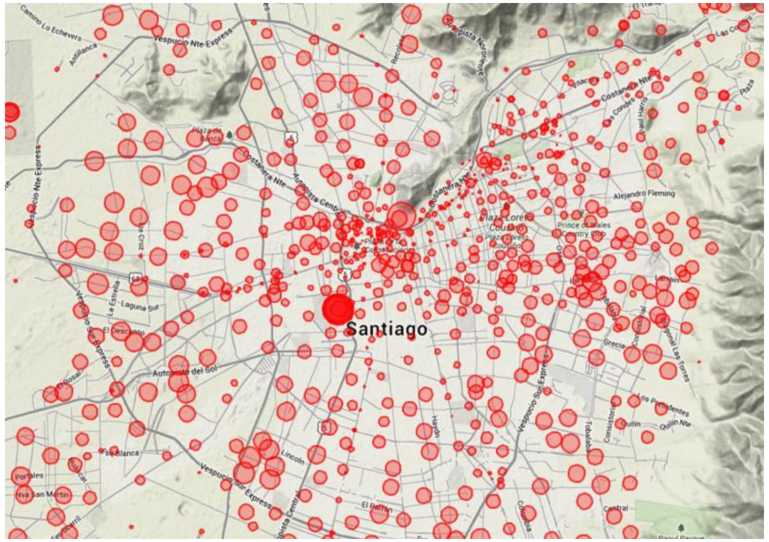


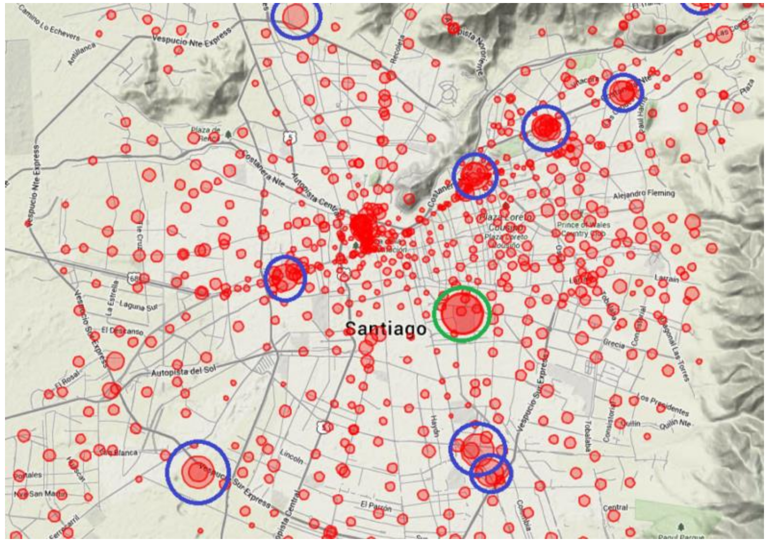
Offices Areas

Experiment 1: A spatial comparison

Geographical representation of human behavioral patterns

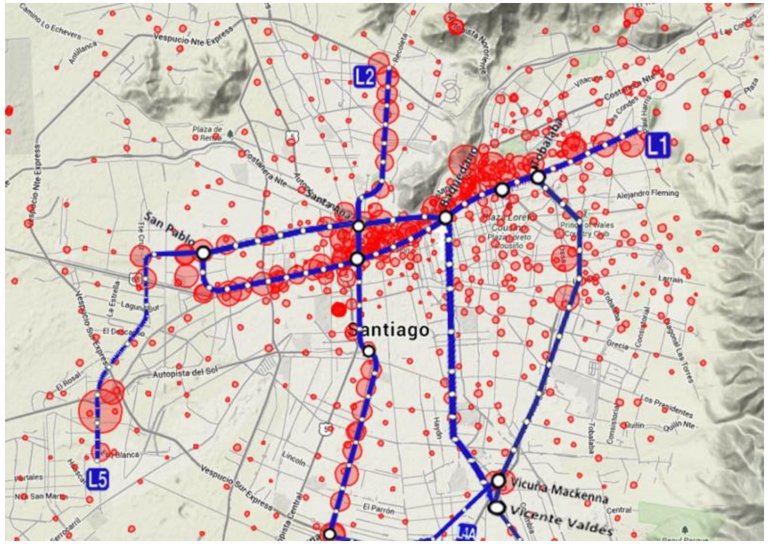
Residential

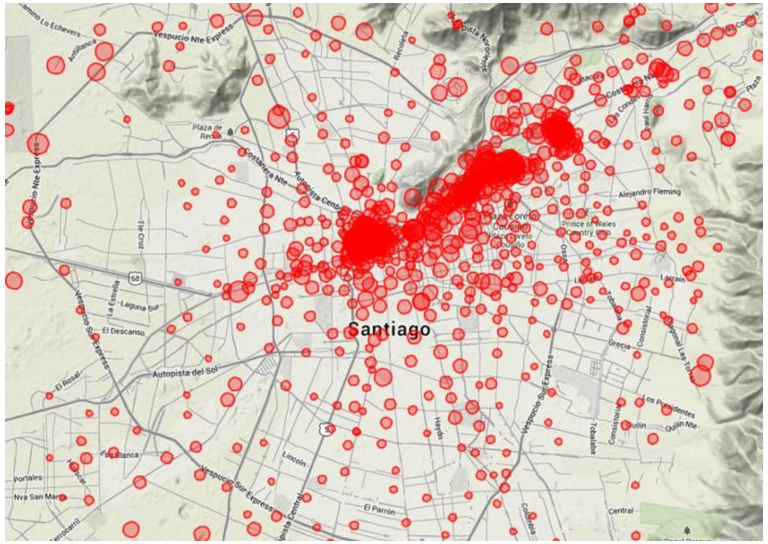




Leisure-Commerce

Rush Hour





Offices Areas

Experiment 1: Conclusions

Spatial human behavioral patterns

We have applied the proposed methodology to understand the behavior of a city by discovering human activity patterns.

- Novel approach using latent variables
- Massive dataset
- We discovered four human behavioral patterns.
- Two of these patterns are very well known (human activity associated with office and residential areas)
- Stability analysis
- Expert knowledge validation

Contents

Experiment 2: Multiple static models

Spatiotemporal human behavioral patterns

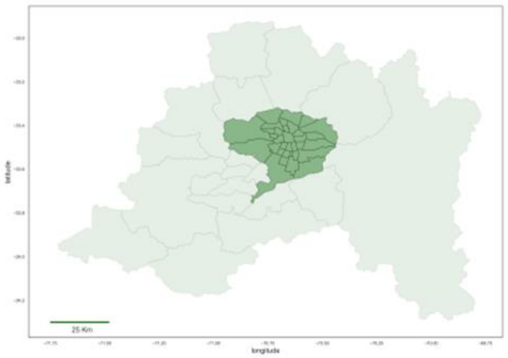
This experiment was designed to address Aim 2 and Aim 3 of this thesis, determine if it can detect human behavioral patterns from digital traces using new data sources, and study their stability. Moreover, this experiment incorporates the temporal dimension in identifying human behavioral patterns.

Banking

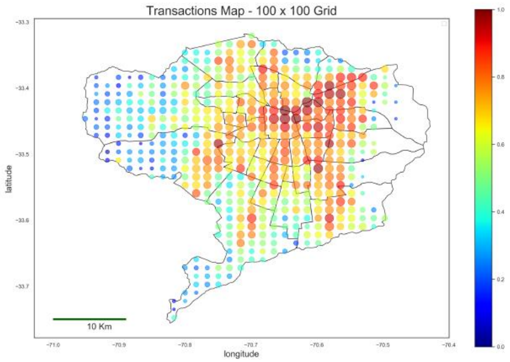
Credit card purchases

190 million records
100,000 terminals
3 years

Dataset



Study Area



Spatial Aggregation

Experiment 2: Experimental Setup

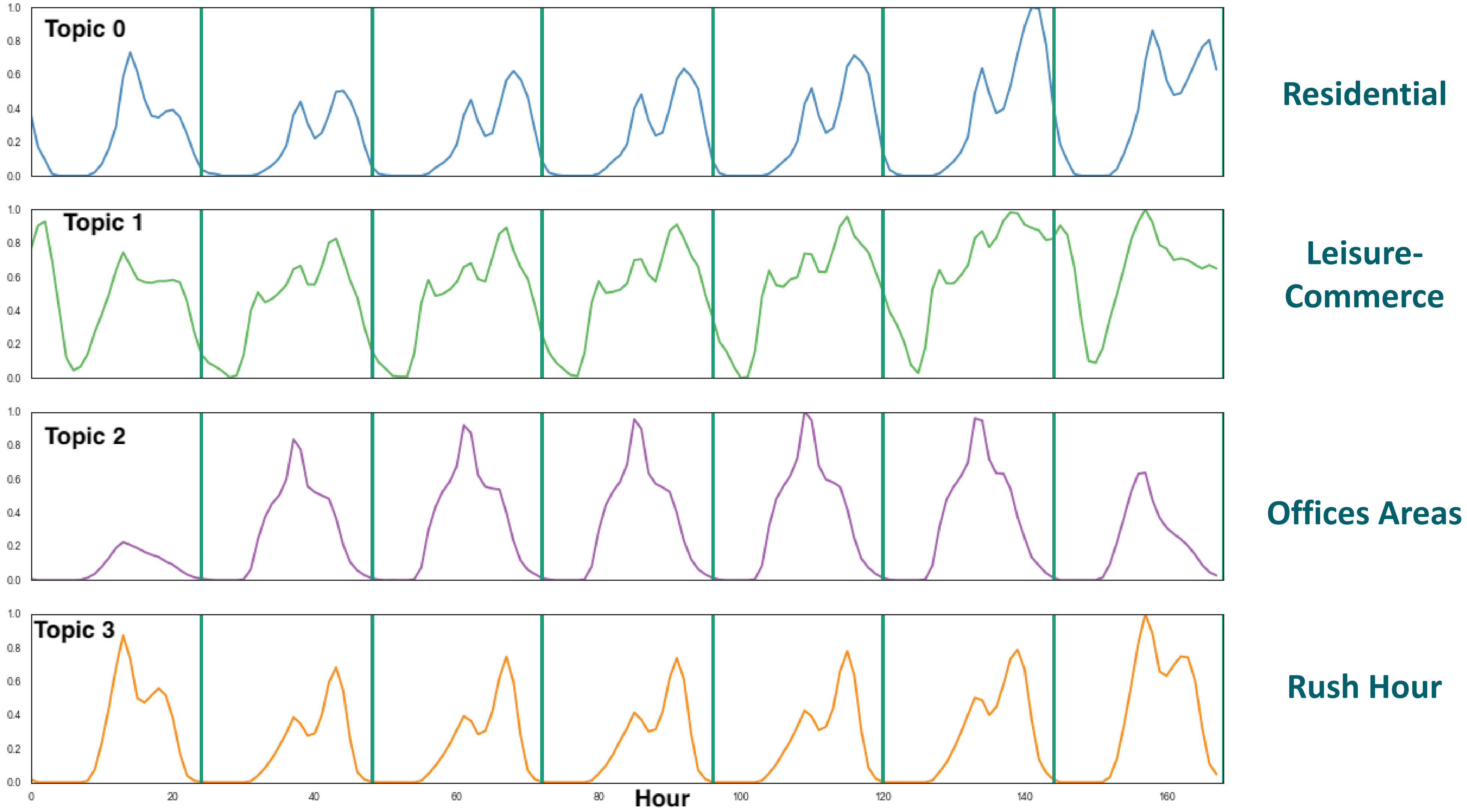
Spatial human activity patterns

We run a series of experiments for each geographical area varying the following parameters:

- **Study Area:** Greater Santiago, Inner Santiago, Downtown Santiago
- **Spatial Aggregation:** 100x100 grid, 400x400 grid
- **Weights:** number of cells (C), number of transactions (T), log(C), log(T)
- **oversampling size:** 10,20,40,60,80,100
- **Models:** LDA, Mini Batch K-Means, Agglomerative Clustering, Gaussian Mixture and Bayesian Gaussian Mixture
- **Number of topics:** 2,3,4,5,6

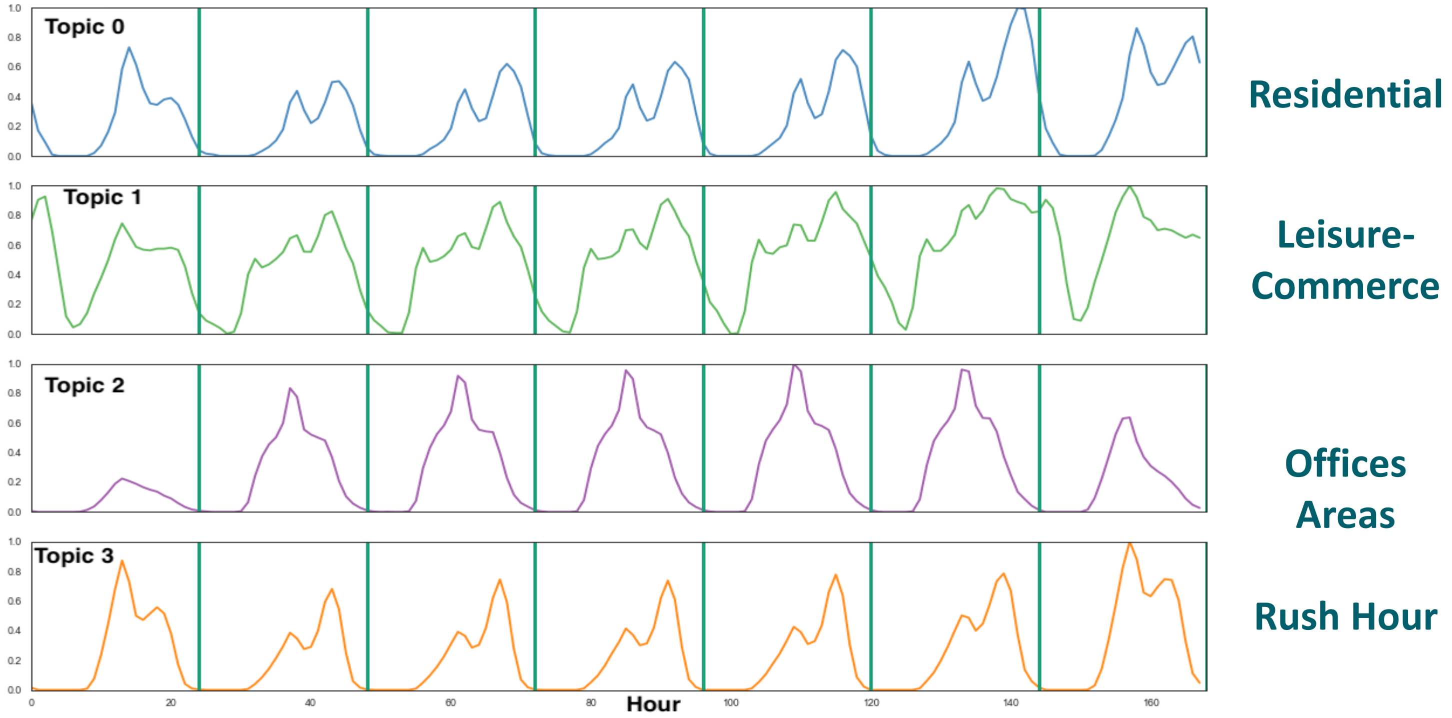
Experiment 2: Results

Spatial human activity patterns



Experiment 2: Results

Spatial human activity patterns



		Telecom dataset			
		T0	T1	T2	T3
Banking dataset	T0	0.63	**0.82**	0.72	0.42
	T1	**0.80**	0.76	0.76	0.70
	T2	0.60	0.41	0.59	**0.93**
	T3	0.69	0.63	**0.94**	0.56

Cosine Similarity between human behavioral patterns discovered using the banking and telecom dataset in the Santiago city area

Experiment 2: Results

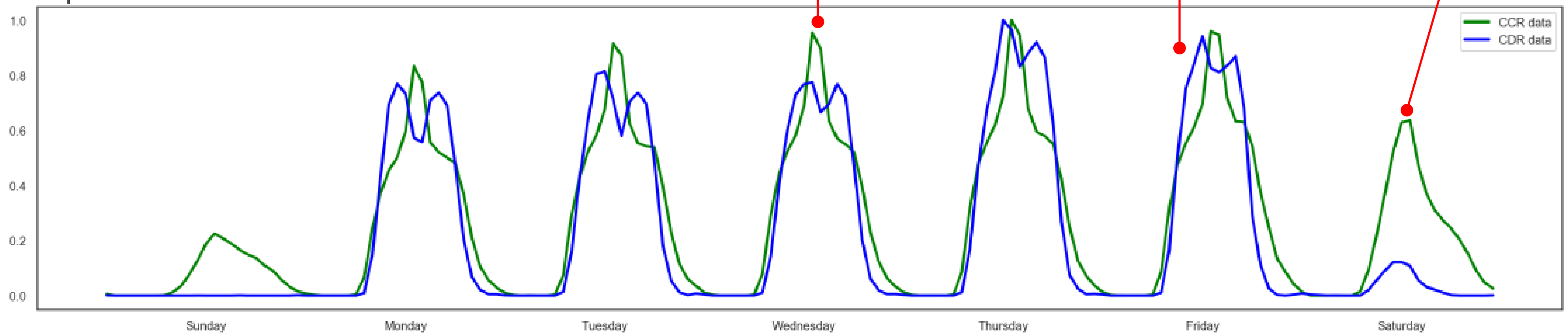
Spatial human activity patterns

Lunchtime

Telecom Dataset (blue line)

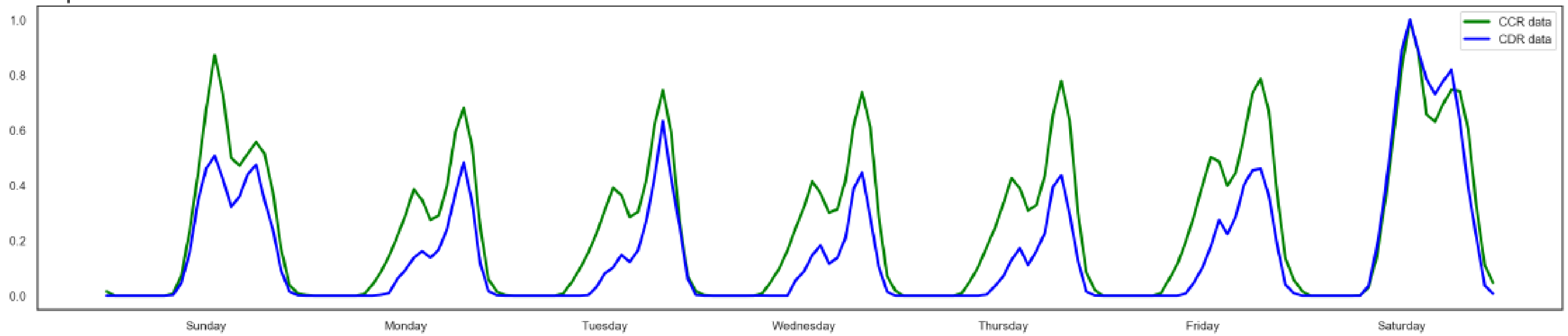
Banking Dataset (green line)

Topic 1



Offices Areas

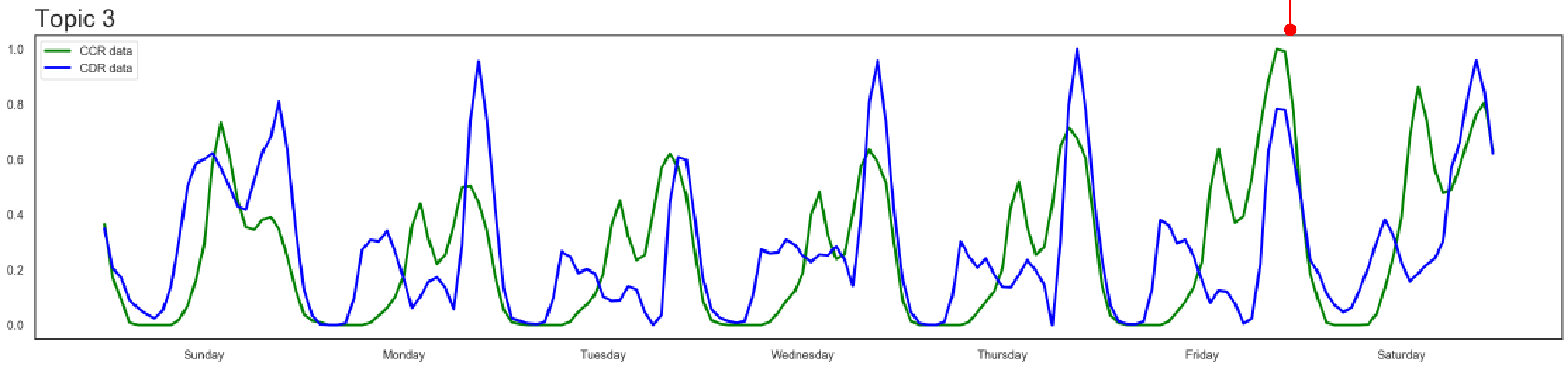
Topic 2



Leisure-Commerce

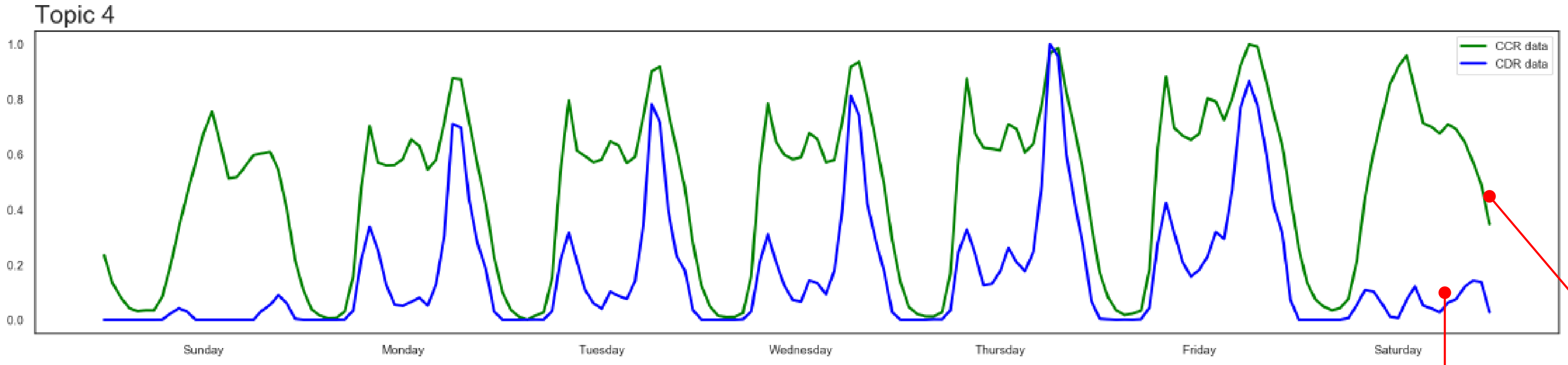
Experiment 2: Results

Spatial human activity patterns



Afternoon peaks (21:00 hrs)

Residential



Rush Hour

Telecom Dataset (blue line)

Banking Dataset (green line)

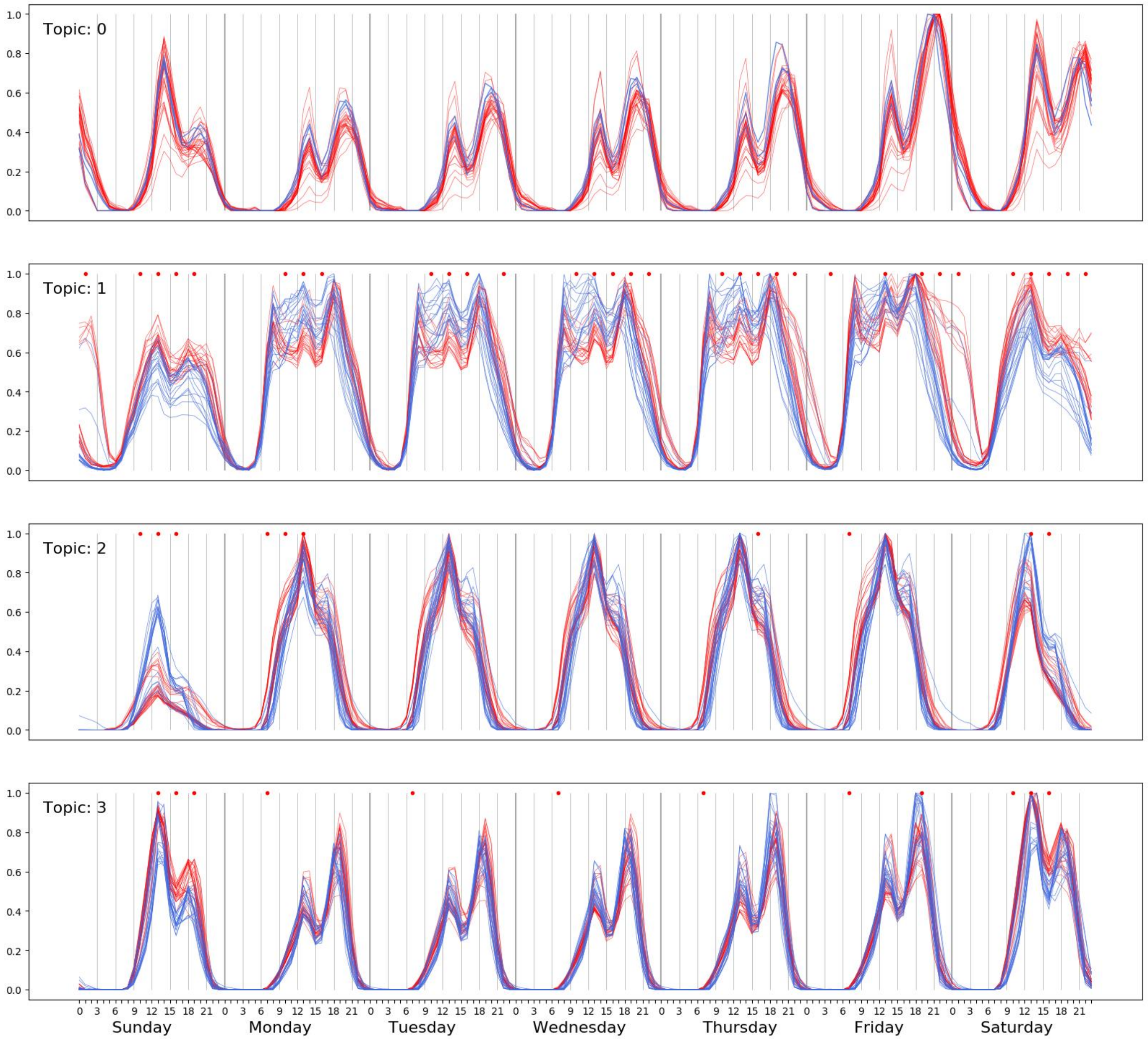
Experiment 2: A spatial comparison

Geographical representation of human behavioral patterns

Spatial distribution of Leisure-Commerce activity topic obtained using the banking dataset, overlaid by the localization of the main shopping malls (Red markers). Blue color blobs spot the localization of POS with a high contribution of this activity topic in their LDA decomposition.

Experiment 2: Spatiotemporal human behavioral patterns

Assessment of the impact of COVID-19 in human behavioral patterns



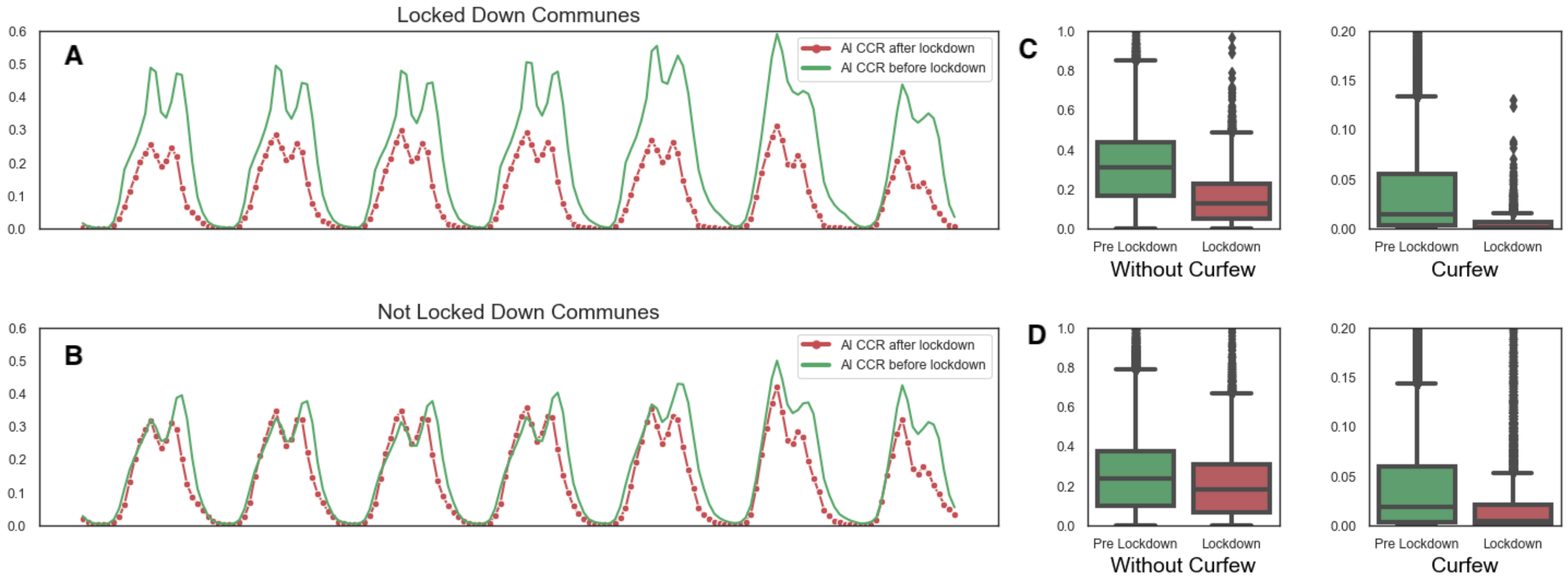
- we collected the credit and debit card transactional data in the **pre-pandemic** and **pandemic** periods.
- we extracted the LDA topics (k = 4) of activity patterns from **windows of 12 consecutive weeks**, with an overlap of 10 weeks between consecutive windows

Change in activity topics due to the lockdowns and curfews imposed to curb the pandemic. Dark blue and red lines correspond to topics extracted from data from 2020 and 2019, respectively.

The red dot denotes statistically significant ($p < 0.0001$) differences among pre-pandemic and pandemic activity topics in aggregations of 3 hours.

Experiment 2: Spatiotemporal human behavioral patterns

Assessment of the impact of COVID-19 in human behavioral patterns



The effect of lockdown and curfew policies in Santiago, Chile.

(A) Average weekly activity before and after lockdown for communes enforcing lockdown

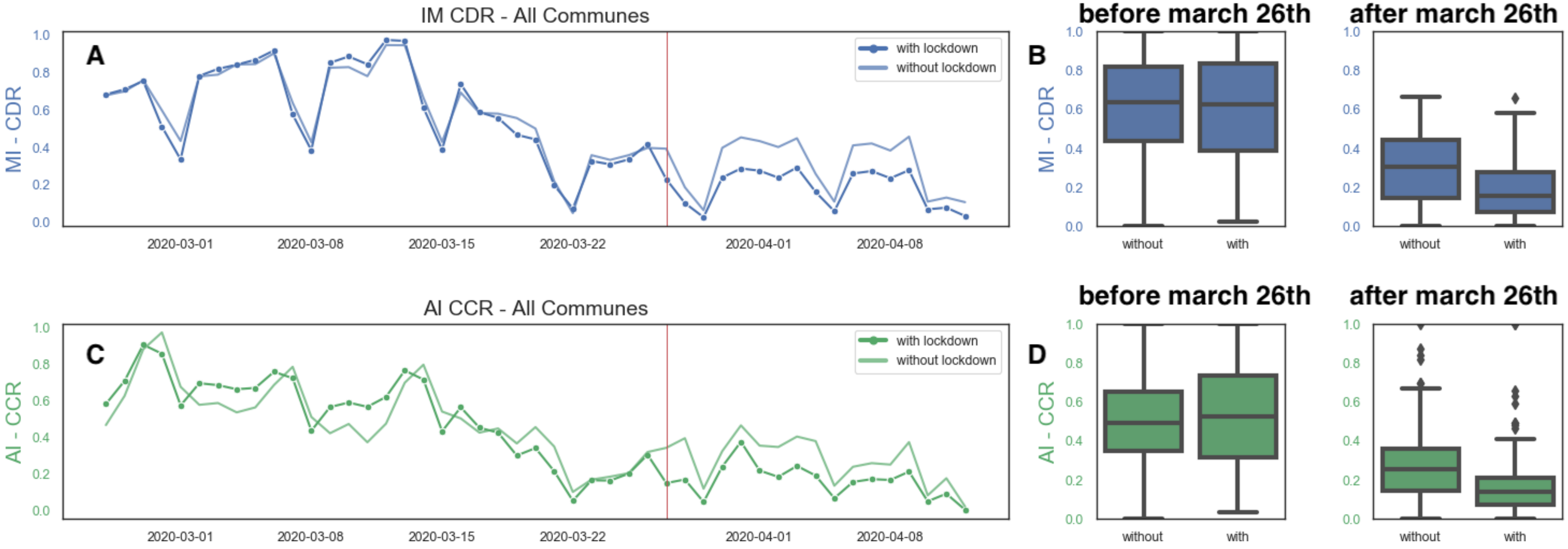
(B) Same for communes not enforcing lockdown

(C) Additional impact of curfew on communes that enforced lockdown

(D) Same for communes that didn't enforce lockdown

Experiment 2: Spatiotemporal human behavioral patterns

Assessment of the impact of COVID-19 in human behavioral patterns



The effect of lockdown policies in Santiago, Chile. Aggregated data from the beginning of March 2020 until April 15th. The red line indicates March 26th.

(A) CDR activity for communes with and without lockdown.
(B) Box-plots of CDR activity in communes with and without lockdown before and after March 26th.
(C) CCR activity for communes with and without lockdown.
(D) Box-plots of CCR activity in communes with and without lockdown before and after March 26th.

Experiment 2: Conclusions

Spatiotemporal human behavioral patterns

We have applied the proposed methodology to understand the behavior of a city by discovering human activity patterns.

- We confirmed the value of using latent variables to detect behavioral patterns
- Massive dataset
- We discovered four human behavioral patterns related to the first experiment patterns
- We included long-term patterns by training multiple models
- More robust stability analysis
- Expert knowledge validation
- We used the proposed methodology to assess the effectiveness of non-pharmaceutical measures to prone the covid-19 effects.

Contents

Experiment 3: Model-embedded patterns

Spatiotemporal human behavioral patterns

This experiment was designed to address Aim 1, Aim 2, and Aim 3 of this thesis, determine if it is possible to detect human behavioral patterns from digital traces using new data sources and models, and study their stability. Moreover, this experiment incorporates the temporal dimension in identifying human behavioral patterns.

Social Media

Geotagged urban activities

32 million records
140 cities
17 years

Dataset

Study Area and Spatial Aggregation

Experiment 3: Experimental Setup

Spatiotemporal human activity patterns

This experiment responds to the three objectives set out in this thesis, **formalizes** the human behavioral pattern validation (Aim 1), **collects and transforms** digital traces to detect behavior patterns (Aim 2), and finally **proposes a method** to identify spatiotemporal patterns in such a way that the temporal evolution of the patterns is captured directly by the proposed model (Aim 3).

- The first step involves comparing results obtained by applying the methodology proposed in this study (DTM) with existing models for detecting activity patterns (DTM, LDA, K-Means, k-Shape, Time series K-Means)
- Temporal matching heuristic
- Time-slices Aggregation
- Model comparison and the optimal number of human behavior patterns
- Final Model
- Human behavior patterns characterization

Experiment 3: Temporal matching heuristic

Spatiotemporal human activity patterns

This section analyzes if it is feasible to align the labeling of human behavior patterns so that similar topics across sequential time slices get the same label.

Time Aggregation	Model	#Topics			
		2	3	4	5
one-year	K-Means	0.97%	7.57%	5.76%	16.46%
	k-Shape	1.09%	2.45%	0.78%	1.70%
	LDA	0.00%	0.00%	0.29%	0.09%
	TS K-Means	0.98%	0.87%	2.62%	2.41%
three-year	K-Means	3.22%	4.33%	3.53%	4.21%
	k-Shape	1.62%	0.05%	0.48%	0.89%
	LDA	0.00%	0.00%	0.35%	0.00%
	TS K-Means	3.67%	2.70%	4.99%	6.52%

Time matching impact is measured using the **Intertemporal Stability** with cosine similarity. The percentual increment between the heuristic arrangement and the original topic labels is shown.

Experiment 3: Time-slices Aggregation

Spatiotemporal human activity patterns

This section compares two time-frame choices to train the activity pattern detection models.

Topics	Model	Intertemporal Stability			Intratemporal Similarity			Topic Smoothness			Topic Consistency		
		3Y	1Y	DIFF	3Y	1Y	DIFF	3Y	1Y	DIFF	3Y	1Y	DIFF
$k=2$	K-Means	0.31	0.31	-0.58%	0.36	0.42	-13.6%	5.25	15.39	-65.8%	0.82	0.75	8.6%
	k-Shape	0.08	0.14	-40.3%	0.06	0.15	-58.5%	1.78	3.44	-48.1%	0.97	0.92	5.0%
	TS K-Means	0.31	0.29	6.5%	0.36	0.39	-8.4%	5.25	11.15	-52.8%	0.82	0.76	7.1%
	LDA	0.07	0.07	-2.42%	0.03	0.08	-60.3%	1.93	2.92	-33.9%	0.96	0.92	4.9%
	DTM	0.00	0.07	-89.3%	0.32	0.43	-24.4%	1.78	3.78	-52.9%	0.98	0.88	11.0%
$k=3$	K-Means	0.29	0.41	-27.4%	0.31	0.48	-35.1%	5.37	18.66	-71.1%	0.81	0.65	24.1%
	k-Shape	0.12	0.14	-15.5%	0.05	0.16	-68.5%	1.82	3.62	-49.7%	0.97	0.89	9.0%
	TS K-Means	0.29	0.38	-21.8%	0.31	0.46	-32.0%	5.37	14.04	-61.6%	0.81	0.67	21.0%
	LDA	0.05	0.12	-60.2%	0.12	0.23	-49.0%	2.51	4.68	-46.2%	0.92	0.85	7.8%
	DTM	0.00	0.09	-93.4%	0.40	0.49	-17.4%	2.00	5.91	-66.0%	0.97	0.77	25.2%
$k=4$	K-Means	0.34	0.43	-20.5%	0.35	0.47	-24.9%	6.47	20.5	-68.4%	0.76	0.65	17.1%
	k-Shape	0.19	0.18	6.3%	0.12	0.15	-21.8%	2.88	3.6	-20.0%	0.91	0.88	3.5%
	TS K-Means	0.34	0.41	-15.5%	0.35	0.46	-22.8%	6.47	15.95	-59.4%	0.76	0.66	15.4%
	LDA	0.04	0.14	-71.8%	0.15	0.26	-42.3%	2.95	5.17	-42.9%	0.89	0.83	7.7%
	DTM	0.00	0.1	-94.7%	0.42	0.52	-18.1%	2.24	6.72	-66.6%	0.96	0.74	30.6%
$k=5$	K-Means	0.35	0.57	-37.4%	0.37	0.61	-38.8%	6.38	25.03	-74.4%	0.73	0.54	35.5%
	k-Shape	0.15	0.24	-34.8%	0.10	0.21	-49.5%	2.68	5.33	-49.7%	0.92	0.85	8.5%
	TS K-Means	0.31	0.46	-32.6%	0.31	0.46	-32.5%	5.64	17.47	-67.6%	0.78	0.65	19.1%
	LDA	0.02	0.14	-84.3%	0.14	0.26	-45.9%	2.88	5.47	-47.2%	0.89	0.83	7.3%
	DTM	0.00	0.09	-95.1%	0.37	0.63	-39.8%	2.58	9.22	-72.0%	0.95	0.71	33.2%

Experiment 3: Model comparison

Model comparison and the optimal number of human behavior patterns

Intertemporal Stability Index using cosine distance

Model	#Topics			
	2	3	4	5
K-Means	0.31	0.29	0.34	0.35
k-Shape	0.08	0.12	0.19	0.15
TS K-Means	0.31	0.29	0.34	0.31
LDA	0.07	0.05	0.04	0.02
DTM	0.00	0.00	0.00	0.00

Intratemporal Similarity Index using cosine distance

Model	#Topics			
	2	3	4	5
K-Means	0.36	0.31	0.35	0.37
k-Shape	0.06	0.05	0.12	0.10
TS K-Means	0.36	0.31	0.35	0.31
LDA	0.03	0.12	0.15	0.14
DTM	0.32	0.40	0.42	0.37

Topic Consistency - Cosine Similarity

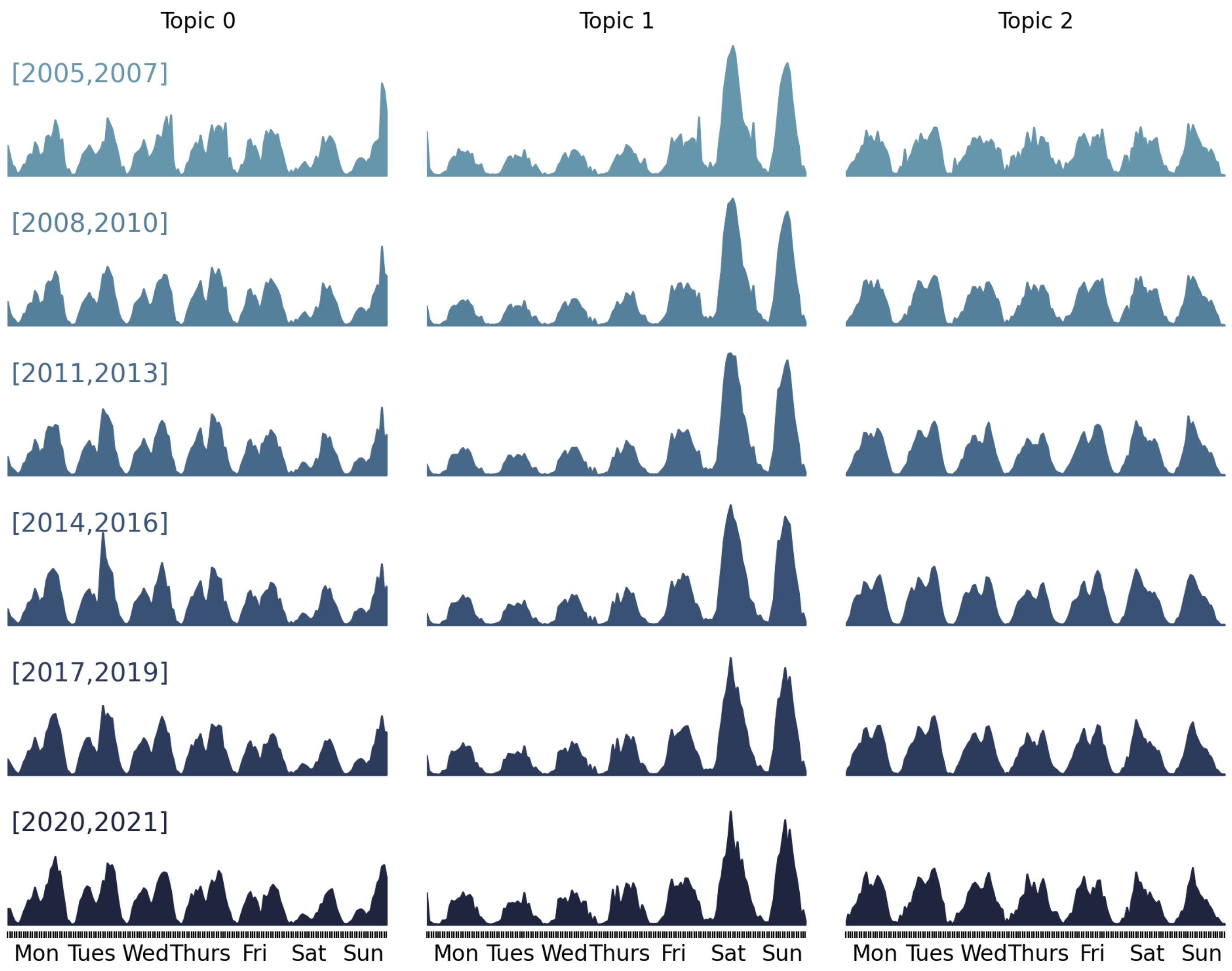
Model	#Topics			
	2	3	4	5
K-Means	0.82	0.81	0.76	0.73
k-Shape	0.97	0.97	0.91	0.92
TS K-Means	0.82	0.81	0.76	0.78
LDA	0.96	0.92	0.89	0.89
DTM	0.98	0.97	0.96	0.95

Topic Smoothness

Model	#Topics			
	2	3	4	5
K-Means	5.25	5.37	6.47	6.38
k-Shape	1.78	1.82	2.88	2.68
TS K-Means	5.25	5.37	6.47	5.64
LDA	1.93	2.51	2.95	2.88
DTM	1.78	2.00	2.24	2.58

Experiment 3: Final Model

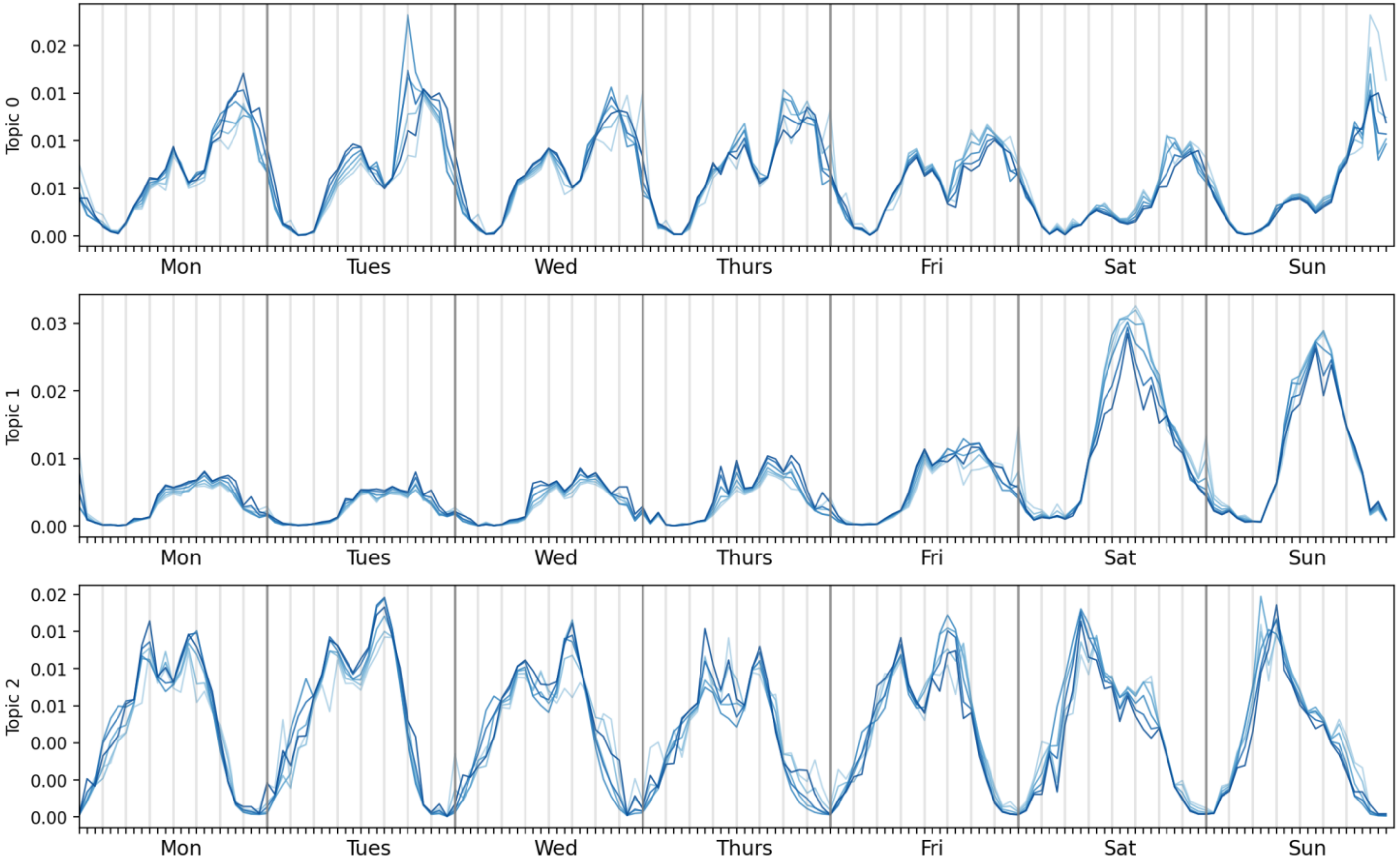
Multi-sensor and multi-temporal city activity patterns from human behavior



- **Topic 0** is characterized by leisure and commerce activities during the week, with two activity peaks at noon and 9 PM. On weekends, the night peak is more significant than the noon peak.
- **Topic 1** has low activity during the week (the activity increases between 9 AM and 9 PM) but significant activity on weekends with a peak at 3 PM that declines abruptly on Sundays.
- **Topic 2** shows a similar pattern during the week and on weekends, with activity concentrated between 9 AM and 6 PM and a slight drop at noon during the week. However, on weekends, activity declines after the 9 AM peak. This pattern is similar to office-areas activity.

Experiment 3: Final Model

Multi-sensor and multi-temporal city activity patterns from human behavior



- **Topic 0** is characterized by leisure and commerce activities during the week, with two activity peaks at noon and 9 PM. On weekends, the night peak is more significant than the noon peak.

- **Topic 1** has low activity during the week (the activity increases between 9 AM and 9 PM) but significant activity on weekends with a peak at 3 PM that declines abruptly on Sundays.

- **Topic 2** shows a similar pattern during the week and on weekends, with activity concentrated between 9 AM and 6 PM and a slight drop at noon during the week. However, on weekends, activity declines after the 9 AM peak. This pattern is similar to office-areas activity.

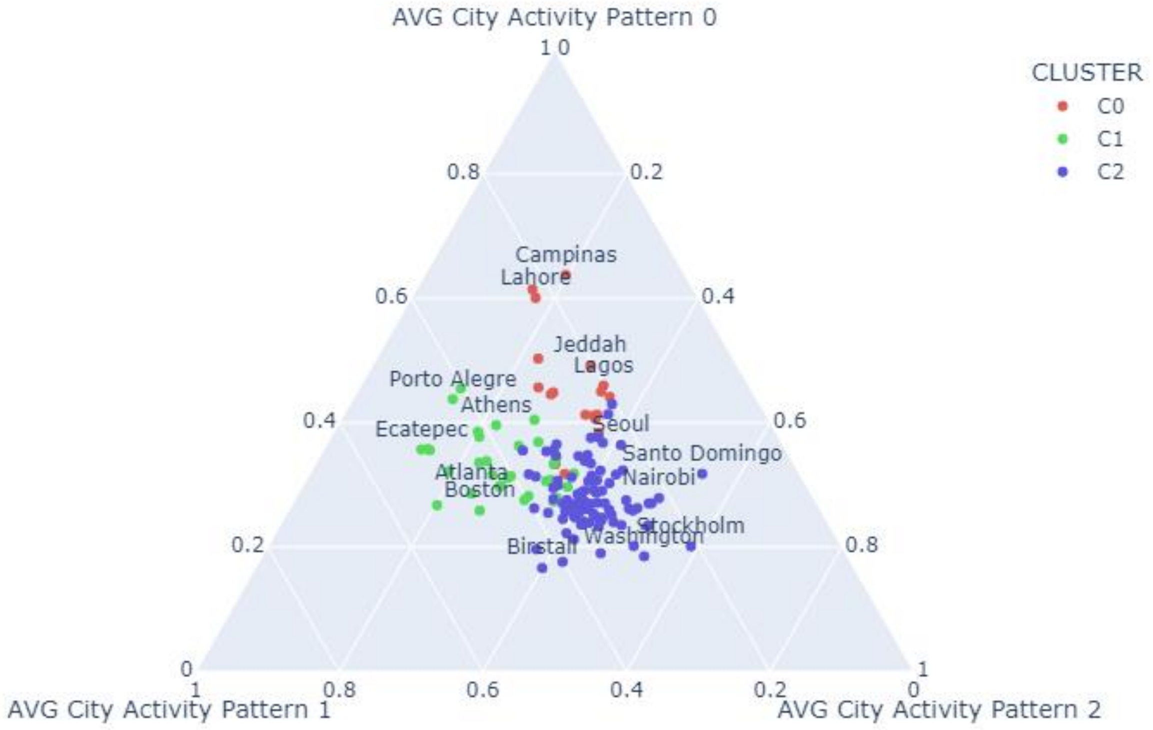
Experiment 3: Characterization

Multi-sensor and multi-temporal city activity patterns from human behavior

After describing each human activity pattern, it is interesting to understand how this information allows us to characterize the behavior of different cities worldwide.

We generated groups of cities based on how Activity Patterns' composition varies over time (mean and std)

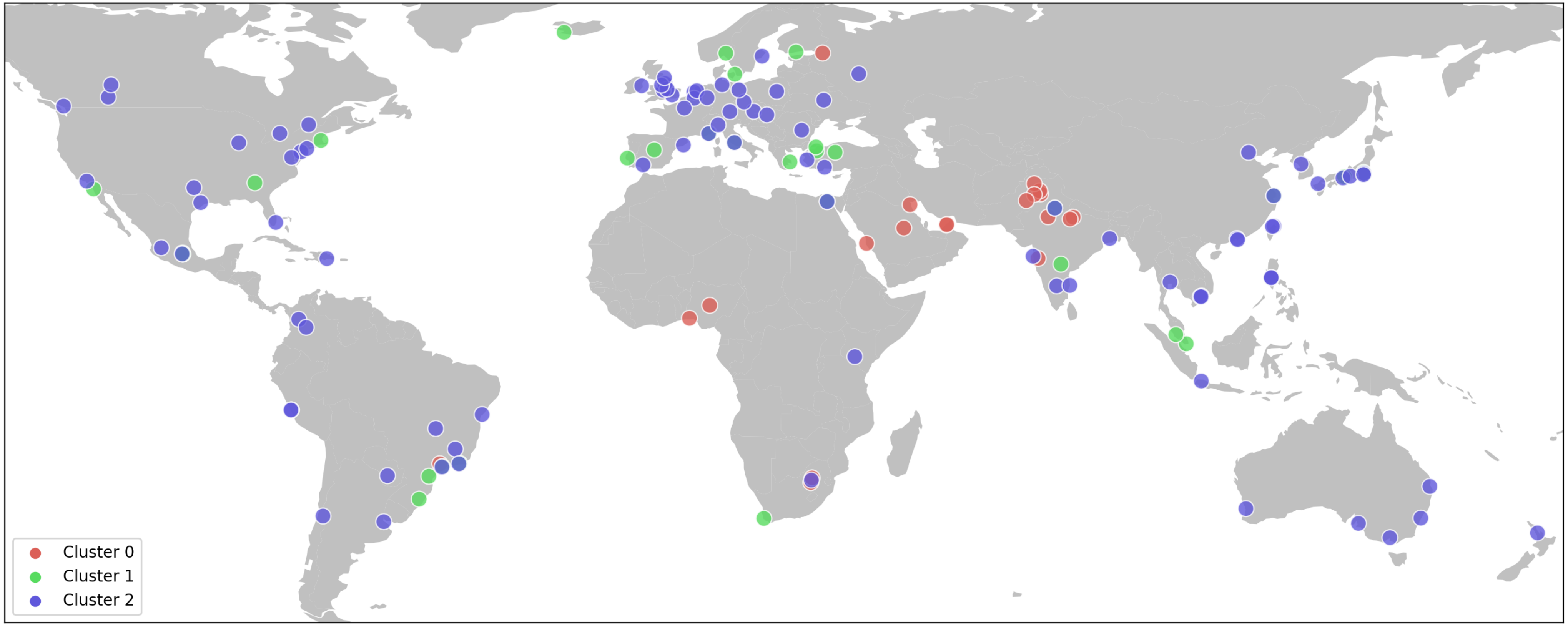
- **C0:** Campinas (Brazil), Lahore (Pakistan), Jeddah (Saudi Arabia), and Lagos (Nigeria)
- **C1:** Porto Alegre (Brazil), Athens (Greece), Ecatepec (Mexico), and Atlanta and Boston (USA).
- **C2:** Seoul (South Korea), Santo Domingo (Dominica Republic), Nairobi (Kenya), Stockholm (Sweden), Washington (USA), and Birstall (England).



Cluster of cities based on their human behavior patterns composition

Experiment 3: Characterization

Multi-sensor and multi-temporal city activity patterns from human behavior



Cluster 0 cities (20) are located mainly in the Middle East, South Asia, and Africa. The cities corresponding to **Cluster 1** and **Cluster 2** are distributed in practically the same territories, except that we did not find any of the 32 cities of Cluster 1 in East Asia and Oceania. Finally, Cluster 2 stands out for having several of its 92 cities in central Europe.

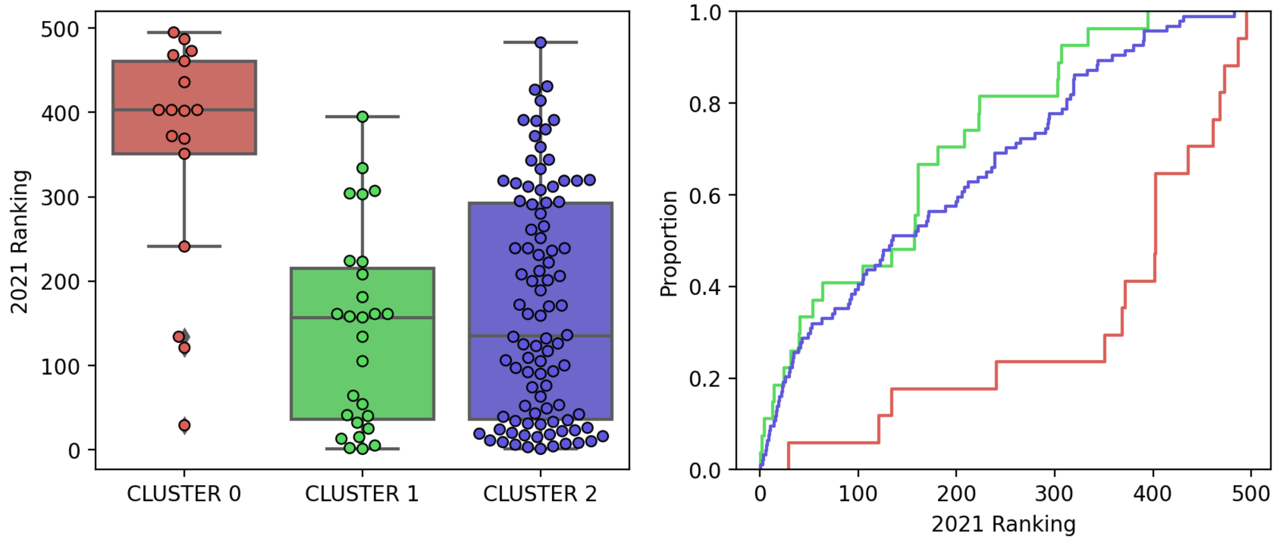
Experiment 3: Characterization

Multi-sensor and multi-temporal city activity patterns from human behavior

The **Innovation Cities Index**, an annual quantitative index to rank the most innovative cities worldwide.

The index is based on cities' **cultural assets**, **human infrastructure**, and **networked markets**.

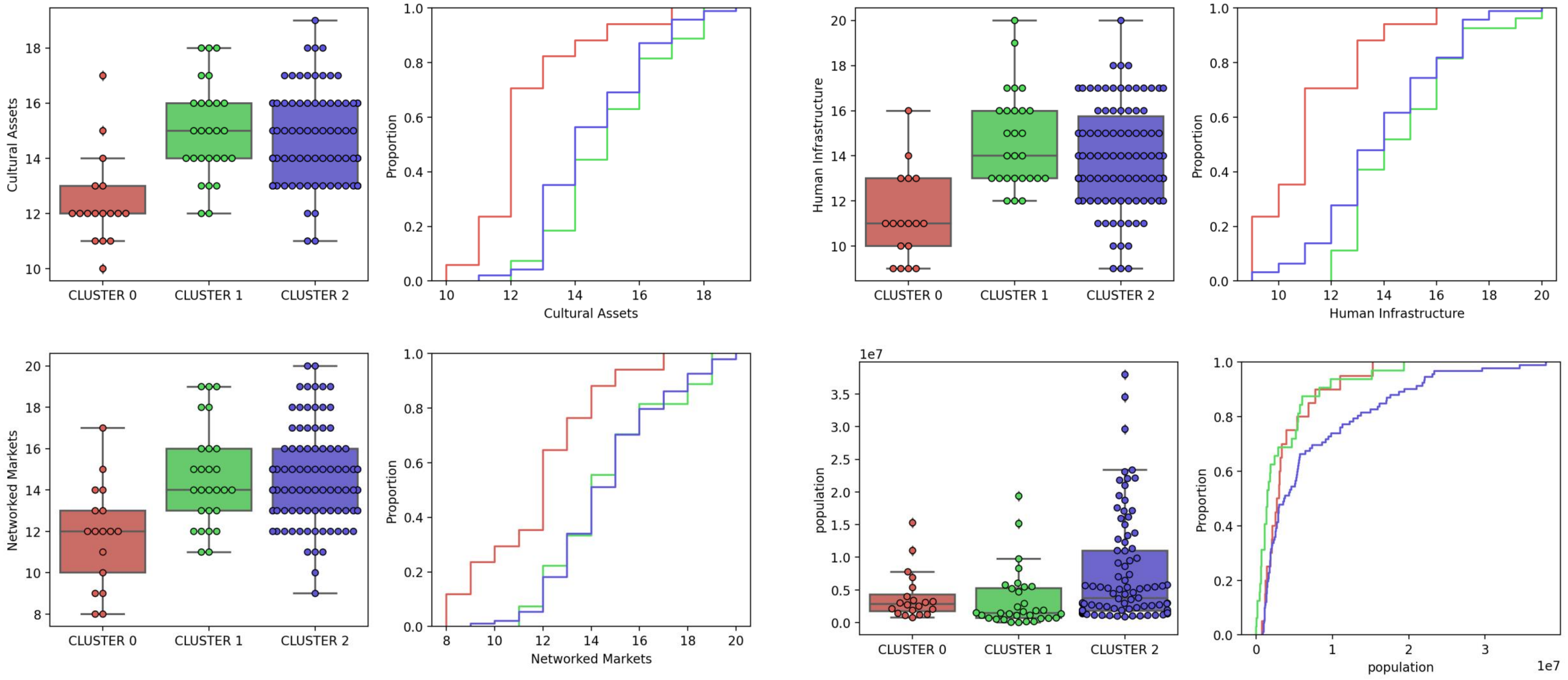
- **Cultural assets** refer to how culture is experienced within cities and considers arts districts, civic institutions, museums, music events, galleries, political protests, books, media, availability of information, and sports.
- **Human Infrastructure** includes the infrastructure deployed in the city for mass transit, finance, universities, hospitals, rail, roads, law, commerce, start-ups, healthcare, and telecommunications.
- **Networked Markets** measure a city's influence and connections in global markets, considering geography, economics, exports and imports, technology, market size, geo-political aspects, and diplomacy.



Innovation Cities Index 2021 by Cluster

Experiment 3: Characterization

Multi-sensor and multi-temporal city activity patterns from human behavior



Experiment 3: Conclusions

Spatiotemporal human behavioral patterns

We have applied the proposed methodology to understand the behavior of a city by discovering human activity patterns.

- Novel approach using dynamic topic models
- Massive dataset
- We included long-term patterns by training a single model
- Reduced expert knowledge validation
- We used the results to characterize the different city behaviors

Contents

Conclusions

On the use of multi-sensor digital traces to discover spatio-temporal human behavioral patterns

- Technology is present in people's lives and leaves **digital traces** of their behavior.
- Georeferenced digital traces of individual behavior help study urban planning, infrastructure management, public transportation, and public policies.
- This thesis addressed the study of **multi-sensor and multi-temporal human behavior patterns** from digital traces with three objectives covered in three extensive studies.

    - The first experiment proposed **alternatives to traditional algorithms**
    - The second experiment addressed identifying spatio-temporal human behavior patterns by training **multiple time-windowed spatial models**.
    - The third experiment **formalizes the validation** of the human behavior pattern through a set of metrics to reduce dependency on extensive expert knowledge of the geographical area studied. Moreover, detect spatio-temporal human behavior patterns **by training a single model.**

Future Work

On the use of multi-sensor digital traces to discover spatio-temporal human behavioral patterns

In this study, **we aggregated spatial information** in different forms, such as Voronoi zones, grids, and cities, for different datasets. Similar behavioral patterns emerged regardless of the aggregation unit and different sensors in different spatial aggregations.

Future research will analyze **the impact of spatial aggregation** on patterns obtained and investigate any **hierarchical dependency** in these aggregations.

A set of metrics was proposed to **reduce dependence on extensive expert knowledge** and formalize the validation of patterns obtained in the last experiment.

Future work will study the **mathematical properties** of these metrics and measure their impact on the shape of the final patterns.

Thank you!

Any Question?

#343838 #005F6B #008C9E #00B4CC #00DFFC