Unveiling social vibrancy in urban spaces with app usage

Thomas Collins*[,1], Diogo Pacheco[1], Riccardo Di Clemente[2,3], and Federico Botta[1,2]

[1] *Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United Kingdom.*
[2] *The Alan Turing Institute, London, NW1 2DB, United Kingdom.*
[3] *Complex Connections Lab, Network Science Institute, Northeastern University London, London, E1W 1LP, United Kingdom.*
*trc207@exeter.ac.uk

Urban vibrancy is an important measure of the energetic nature of a city that is related to why and how people use urban spaces, and it is inherently connected with our social behaviour. Increasingly, people use a wide range of mobile phone apps in their daily lives to connect socially, search for information, make decisions, and arrange travel, amongst many other reasons. However, it is not clear whether the usage of such apps is influenced or related to the urban spaces in which people are, and the demographics of those users. Here, we use app-usage data as a digital signature to investigate this question. To do this, we use the `NetMob23` data set, a rich data source of spatiotemporal app usage across the three largest cities in France. We frame the analysis within the fine-grade *IRIS2000* subdivisions using both urban features (constructed from *OpenStreetMap* data) and socioeconomic indicators (constructed from *IRIS2000* data). We develop a series of univariate spatial models and interpret the results alongside an HDBSCAN clustering approach. Our analysis finds evidence for work, education, and caregiving behaviours. We found differences across cities and between the week and the weekend. Our results add further evidence for the importance of using computational approaches for understanding urban environments, the use of sociological concepts in computational social science and for understanding urban vibrancy in cities.

Keywords: (1) Urban vibrancy, (2) Mobile app activity, (3) Spatial patterns (4) Urban spaces, (5) Social fabric.

1 Introduction

Cities are one of the largest and most complex man-made systems [1]. They are constantly growing and transforming, placing increasing pressure on them to facilitate more people with increasing opportunities. Whilst cities can generate wealth, they might also create or increase inequalities. Researchers and urban planners have traditionally relied on low-frequency data gathering to measure our cities [2]. Subsequently, hidden inequalities, which cannot be easily addressed by policymakers or urban planners, can become deeply ingrained, resulting in unequal access to opportunities or city infrastructure. These problems can be related to the physical urban environments of cities, such as the presence or absence of *'Points of Interest'*, or the underlying socioeconomics of cities, such as the resident's levels of education or population dynamics and structure [3].

Thanks to the recent explosion in the availability of digital forms of data [4], we are now able to measure cities at a higher frequency than ever before [5]. Coupled with advanced computational methods [6], this relatively new ability to track and understand peoples' mobility allows us to understand social behaviour in terms of inequalities stemming from phenomena such as socio-spatial segregation [7]. Furthermore, because aggregated data can mask inequalities, research has attempted to reveal inequalities by investigating highly granular data sets and their variation across space, time, and socio-demographics [8].

This also holds for the study of *Urban vibrancy* (or urban vitality), which refers to the liveliness and energy of a place in an urban context [9, 10, 11]. Urban vibrancy relates to many social

aspects of city life and can be used to understand why and how people use urban space. To measure urban vibrancy, it is becoming increasingly common to use mobile phone data because of the deep interaction that people have with technology and how it has integrated into people's lives via work, home, or social aspects. Urban features are important when considering urban vibrancy because they are related to urban activity [12, 13, 10, 14]. Buildings, highways, and 'Points of Interest' can provide opportunities, housing jobs or offer activity spaces for social meetings. It is understood that both the density and diversity in urban features are particularly influential and this is thought to be because a larger variety of people are drawn to frequent those locations [15].

This study follows on from work by [10] and [14] on urban vibrancy in cities that found evidence for the importance of *third places* whilst using data across social groups to investigate potential spatiotemporal segregation in cities. We extend these ideas by using the `NetMob23` data set, a rich source of spatiotemporal *mobile service usage* ('app-usage') data from *Orange* [16]. Here, we explore the relationship between our usage of apps (across a wide range of app categories), the urban environments in which we use them, and the demographic of the population living in cities. We focus on the three largest cities of France: Paris, Marseille, and Lyon. From the `NetMob23` data set, we use each mobile application consumption volume reduced to their Apple Store group. We consider the clustered nature of app usage and subsequently, to what degree app usage has a spatial component, and how this relates to both urban features and socioeconomics. We use a series of univariate models as a modelling approach because of the richness and multifaceted nature of the data. We use a computational approach for these research aims: (1) we use an HDBSCAN clustering approach to explore the app-usage clusters and how they relate to urban environments and demographic groups, and (2) we use a series of univariate spatial regression models to explore the spatial components of the app usage.

2 Data

We use three main data sources: (1) the `NetMob23` data set which contains the usage of mobile applications over time and across 20 cities of France, (2) the *IRIS2000* data for population-level and education-level information of the residents of the cities in the analysis, (3) `OpenStreetMap` data [17]–a crowdsourced data used and built by the communities that make up the cities themselves. For this analysis, we focus on the cities, and the metropolitan areas of Paris, Marseille, and Lyon: the three largest cities in France.

2.1 `NetMob23` mobile network traffic

We retrieved mobile network traffic data for the areas taken from the `NetMob23` data repository. The data consists of 77 days from the 16$^{\text{th}}$ of March 2019, to the 31$^{\text{st}}$ of May 2019. The data are the up-link and down-link of the volume of demand for 68 mobile services. The data are available at the temporal granularity of fifteen-minute intervals. The values have undergone a normalisation that conceals sensitive information whilst preserving the signal in the data. This means that there are no units of measurement. At download, coverage is represented as probabilities of association and is partitioned into tiles measuring $100 \times 100\,\text{m}^2$ each. For further details on the generation of the data set, see [16]. From this data, we create digital signatures derived from total app usage to investigate digital signatures regularities in these, and to better understand variations across space, time, and sociodemographics.

2.2 *IRIS2000*

IRIS2000 is a division of the geography of France into approximately equally sized areas ('cells'). The system is used in the French census and for population-level statistics carried out by the National Institute of Statistics and Economic Studies (INSEE). The cells have a target size of 2,000

residents. Across France, there are approximately 16,100 IRIS cells. The data are open-source and accessible via download [18]. From this data, we calculate both education– and population–level socioeconomic indicators (explained in further detail in Section 3.2).

2.3 *OpenStreetMap* data

We use *OpenStreetMap* [17] to construct urban features for our analysis. *OpenStreetMap* is a data repository built and maintained by crowdsourcing and collaboration. Volunteer users collect geospatial data and upload it to the open-source repository for collecting and storing digitised representations of urban environments. Subsequently, we retrieve data for a range of urban attributes such as transportation networks or 'Points of Interest' [19]. We downloaded the most up-to-date data for all *IRIS2000* cells in the analysis. From this data, we calculate urban feature metrics (explained in further detail in Section 3.2).

3 Methods

3.1 Digital signatures

We study how usage of a wide range of mobile applications varies by urban areas and socio-demographic groups by creating *digital signatures* from the mobile network traffic data. We aggregate the data by day of the week into two categories: 'Weekdays,' which includes all data from Monday to Thursday, and 'Weekends,' which includes all data from Friday to Sunday. This aggregation allows us to study the emergence of behavioural differences between weekdays, typically reserved for work and study, and weekends, where social and leisure activities usually take place. We then use areal interpolation [20, 21] to spatially aggregate these data from the original $100 \times 100\,\mathrm{m}^2$, lowering the resolution to match that of the *IRIS2000* cells. This allows us to investigate the relationship between app usage and socio-demographic groups.

We group each of the 68 mobile services in the `NetMob23` data set according to the *Apple Store* app categorisation; these are: 'Adult Content', 'Advertising', 'App Store', 'Cloud Services', 'Cloud Storage', 'E-Commerce', 'Email', 'File Sharing', 'Finance', 'Food', 'Gaming', 'Live Streaming', 'Maps', 'Messaging', 'Music', 'Privacy', 'Productivity', 'Professional Networking', 'Reference', 'Remote Desktop', 'Shopping', 'Social Networking', 'Streaming', 'Transportation', 'Video', 'Video Conferencing', 'Video Sharing', 'Virtual Assistant', 'Weather', and 'Web Services'. For ease of presentation, in the following, we refer to each app category simply as 'apps', to simplify language, but it is always intended as the aggregation of all apps belonging to a specific category.

We investigate the correlation between the up-link and down-link for each app category. We find that the up-link is correlated with the down-link (all correlations were larger than 0.35, with the mean correlation value larger than 0.8; see Table 1), so we aggregate the down-link and up-link giving the total app usage for each app (n=34; see Figure SI 1 and SI 2). We refer to the collection of apps usage in each cell as the *digital signature* of that cell. For each cell in the *IRIS2000* data, we standardise the digital signature by subtracting the mean and dividing by the standard deviation. This allows us to remove differences between cells with high and low apps usage, focusing instead on how the usage is divided across the different apps, irrespective of the total volume.

3.2 Social and geographical features

3.2.1 Urban features

We download urban feature data for all cities in the analysis from the crowdsourced geospatial data repository: *OpenStreetMap* ('OSM') [17]. Of these geospatial data, we only retrieve features with the following OSM key categories: 'key:amenity'–detailing essential and significant amenities

Table 1: `NetMob23` down-link is correlated with the up-link. The table summarises descriptive statistics (central tendency, dispersion, and shape) for each city, and for each period: 'week' and 'weekend'. These data are the down-link and up-link of the `NetMob23` data set[16]. The values have undergone a normalisation that conceals sensitive information whilst preserving the signal in the data. This means that there are no units of measurement. The minimum was 0.35 (Lyon:Weekend) , however, all mean correlation values were larger than 0.8.

	Paris		**Marseille**		**Lyon**	
	Week	Weekend	Week	Weekend	Week	Weekend
	τ	τ	τ	τ	τ	τ
Count	68	68	68	68	52	49
Mean	0.89	0.88	0.88	0.87	0.88	0.84
Std	0.09	0.1	0.09	0.09	0.1	0.15
Min	0.64	0.58	0.55	0.6	0.53	0.35
25%	0.85	0.84	0.83	0.82	0.84	0.79
50%	0.91	0.9	0.91	0.91	0.92	0.9
75%	0.95	0.95	0.95	0.94	0.95	0.94
Max	0.99	0.99	0.98	0.98	0.99	0.98

catering to both visitors and residents, including restroom facilities, public phones, banking institutions, pharmacies, correctional facilities, and educational institutions; 'key:leisure'–which are mainly locations where individuals frequent during their leisure hours; 'key:shop'–which are for shops; and 'key:sport'–for identifying those locations where sport can be carried out. To improve the signal and focus the analysis on the most relevant 'Points of Interest', we reduce the data set by discounting those POIs that occurred less than 10 times in the data.

We manually label these data using two classification systems: (1) 'Points of Interest' known to be impactful on segregation, and (2) a socially-focused system for '*third places*' (explained below). The first classification system is consistent with that employed by [22] and [23]. This system aligns with the *Foursquare* classification system, encompassing 14 distinct categories: (1) Arts / Museum, (2) City / Outdoors, (3) Coffee / Tea, (4) College, (5) Entertainment, (6) Food, (7) Grocery, (8) Health, (9) Residential, (10) Service, (11) Shopping, (12) Sports, (13) Transportation, and (14) Work. We chose these classification labels because they represent frequently visited locations and are likely to have significance in the context of urban segregation.

We use the labelling methodology as above but follow the concept introduced by [24, 25]. Third Places are distinct from homes or workplaces and hold significant importance in enhancing urban vibrancy [25]. They facilitate spontaneous gatherings in urban settings, contributing positively to communities [10]. Given that people invest a substantial portion of their leisure time in these places, they are a key focus of urban analysis. Third places fall under five categories: (1) eating and drinking, (2) organized activities, (3) outdoor, and (4) commercial venues, and (5) commercial services. Following classification, there were 110 unique third places and 244 unique 'Points of Interest'. With these classification systems, we process both third places and 'Points of Interest' data using two calculations:

(1, 2) Urban feature density (both 'Points of Interest' and 'third places'): defined as the density of the labelled features divided by the area of the cell. Increasing the density of features has been found to increase urban vibrancy [10, 15].

(3, 4) Urban feature diversity (both 'Points of Interest' and 'third places'): defined as the diversity of the labelled features. Again, increasing the diversity of features has been found to increase urban vibrancy [10, 15]. The Shannon-Wiener Index (H) is calculated using the following formula [26]:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i \tag{1}$$

Where:

$$\begin{aligned} H &= \text{Shannon-Wiener Index} \\ M &= \text{Number of different categories} \\ P_i &= \text{Proportion of individuals in category } i \\ \log_2 &= \text{Base-2 logarithm} \end{aligned}$$

3.2.2 Socioeconomic indicators

Population-level Using the *IRIS2000* population-level data set, we calculated a population-level socioeconomic indicator: (5) the dependency ratio.

(5) Dependency Ratio: This is calculated as the ratio of the dependent population of both the youth (under 18) and the elderly (over 60) combined divided by the working-age population (19–59). We use this to assess any economic burden of supporting dependents on the working-age population. We converted these data into feature density estimates. We calculated the dependency ratio using the following formula:

$$\text{Dependency Ratio} = \frac{\text{Number of Dependents}}{\text{Working-Age Population}} \times 100 \tag{2}$$

Where:

$$\begin{aligned} \text{Dependency Ratio} &= \text{Dependency Ratio (in percentage)} \\ \text{Number of Dependents} &= \text{Total number of individuals who are dependents} \\ \text{Working-Age Population} &= \text{Total number of individuals of working age} \end{aligned}$$

Education-level We extended our population-level socioeconomics with education-level indicators. These include: (6) total enrollment, and (7) total out-of-school.

(6) Total enrollment: we also retrieved the total number of people enrolled in education by summing the number of people in education from [2–5] years, [6–10] years old, [11–14], [15–17], [18–24], [25–29], and aged 30 or over. We converted these data into feature density estimates.

(7) Total out-of-education: We also retrieve the total number of people out-of-school who are aged 15 or over. We converted these data into feature density estimates.

3.3 Statistical approach

3.3.1 Clustering approach

We processed the `NetMob23` mobile network traffic data in an HDBSCAN (hierarchical density-based spatial clustering of applications with noise [27]) clustering algorithm such that we could recognise distinct clusters of app-usage. We included the combined categorised app usage as a feature array for our clustering. We included an encoded 'week type' label such that we could compare the week and weekend values separately. It is important to note that *IRIS2000* spatial cells were removed for this analysis so that they wouldn't enforce any spatial proximity on the

clusters. We assess the optimal number of clusters using the silhouette method [28], a technique that measures how well separated the clusters are with respect to the other clusters. The silhouette method uses coefficients for the points measuring the similarity to its own cluster compared to others. For HDBSCAN, this means that we gain the average silhouette score over a varying number of clusters. The highest score is then chosen as the starting value for HDBSCAN. We ran the clustering algorithm using ten runs with different centroids and kept the best of these in terms of inertia.

3.3.2 Spatial modelling approach

Our approach to understanding the relationship between urban vibrancy, urban features, and socioeconomics uses the *IRIS2000* cells. We use those geometries as a reference with our urban features data and socioeconomic indicators, and we interpolated the `NetMob23` apps data using areal interpolation [20, 21].

Our aim is to model each of the app categories for each city while including an aggregated model for clarity and to highlight common trends across the cities in the analysis. This aggregated model will be referred to regularly throughout the analysis. Prior to the analysis, we carried out standardisation of all variables to ensure comparability.

We outline the method for spatial regression here. For each of the dependent variables in the multivariate analysis, we create an ordinary least squares (OLS) as a baseline that we use to evaluate spatial dependence. For this, we use Moran's I analysis of the residuals of the OLS model. This provides statistics pertaining to the existence of spatial structure in the data. Finding spatial structure, and so spatial dependence, would indicate that spatial models were necessary. For spatial modelling, we consider the spatial error model (SER) and spatial lag models (SAR) [29, 30]. These are regression models that account for the spatial component in the data albeit in different ways: SER models incorporate the spatial component within the error term of the regression equation; the SAR includes a spatially lagged variable within the collection of predictors. We calculate the Lagrange multiplier (both non-robust and robust) as well as the Akaike information criterion (AIC). We use these to decide between SAR and SER models and to evaluate the performance of the models. As a spatial weights matrix, we use Queen's contiguity. We construct spatial lag and spatial error models fitting the models with maximum likelihood estimation as outlined in previous work [31, 29]. These models incorporate exogenous variables and provide insights into both direct and indirect effects. We calculated the *total effects*, that is the estimated direct impacts of these variables, which represent the *direct effects*. Additionally, we derived the *indirect effects* by subtracting β (beta) divided by (1 minus the spatial lag term, multiplied by the maximum eigenvalue of the spatial weights matrix) from the direct effects. This calculation enhances our ability to interpret changes in predictor units, considering the presence of spatial spillover effects [32].

Since we are interested in the relationship between urban features, socioeconomics, and urban vibrancy, we included urban features and socioeconomic indicators as independent variables in the model (see Section 3). We separated our modelling to allow spatial and temporal comparisons: (1) models were partitioned by time periods for the weekdays (all data on days from Monday to Thursday) and weekends (all data on days from Friday to Sunday). We did this to maintain separation between the contrasting behaviours between these periods; (2) we used each city in the analysis, and we aggregated data from all cities to observe the combined effects.

4 Results

We use HDBSCAN clustering to investigate the presence of clustering in app usage for each city (Figure 1). In the analysis, each of the three cities had observable clusters for both the week and the weekend: Paris contained five clusters for the week and five for the weekend. Marseille

Table 2: Clustering analysis: descriptive statistics for each city, in each period, and under each label. The table shows the number of *IRIS2000* cells (sum) and the total area (km^2.) in each cluster [18]. These data are following preprocessing taken from the `NetMob23` data set [16]. Note that, for Paris, cluster one and cluster three are shared but that cluster one on the weekend is only based on four cells; this is important when considering the histograms below.

City	Temporal aggregation	labels	No. Tracts	Area
Paris	**Week**	0	909	244.15
	Week	1	375	66.04
	Week	3	28	4.55
	Week	4	641	113.99
	Week	6	965	415.53
	Weekend	1	4	0.29
	Weekend	2	1136	271.31
	Weekend	3	28	4.55
	Weekend	5	1107	473.32
	Weekend	7	643	94.8
Marseille	**Week**	1	397	236.78
	Weekend	0	178	94.19
	Weekend	2	219	142.6
Lyon	**Week**	1	319	368.94
	Week	2	190	154.46
	Weekend	0	509	523.4

contained one cluster for the week and two for the weekend. Lyon contained two clusters for the week but only one for the weekend. Paris had by far the largest amount of cells and area (see Table 2). Paris also had the most clusters. We focus on Paris for the remaining part of this analysis. The clusters of Paris were quite different for the week and weekend, but they shared a few clusters (cluster one and cluster three were shared). Cluster six was the largest. Cluster one made up a large section of the city during the week but diminished during the weekend.

We used density histograms to illustrate the variations in clusters for each variable in the analysis (see Figure 2). Generally, the clusters are similar between the week and weekend. This could be accountable to our 'week type' label which we used to ensure comparison. Despite this reasonable similarity, there are some notable differences. Specifically, for both education variables, clusters one and five show similar shapes with the highest peaks for the weekend. However, this similarity is not observed during the week. In the case of the education enrollment density variable, cluster zero exhibits the highest peak, as shown in Figure 1. The population dependency ratio clusters are more spread for the weekend compared to the week. This could be attributed to week and weekend behaviours. The points of interest diversity show more right skew for the week compared to the weekend. This could be attributed to work-type behaviours where people are using specific locations close to 'Points of Interest'.

We also carried out a series of univariate spatial regression for each of the separated features such that we could pick out further app-specific details. Prior to our spatial modelling approach, we used the residuals of an ordinary least squares (OLS) model within Moran's to test for spatial autocorrelation in the data. This involves determining the global Moran's I value for each city and for each period in the analysis: 'week' and 'weekend'. With these statistics, we can gain an understanding of any spatial dependence in the data. We calculate Moran's I for statistics and a p-value for each of the dependent variables in the spatial modelling analysis. For brevity, we report here the maximum and minimum values for each of the cities in the analysis and at each temporal level (see Table 3).

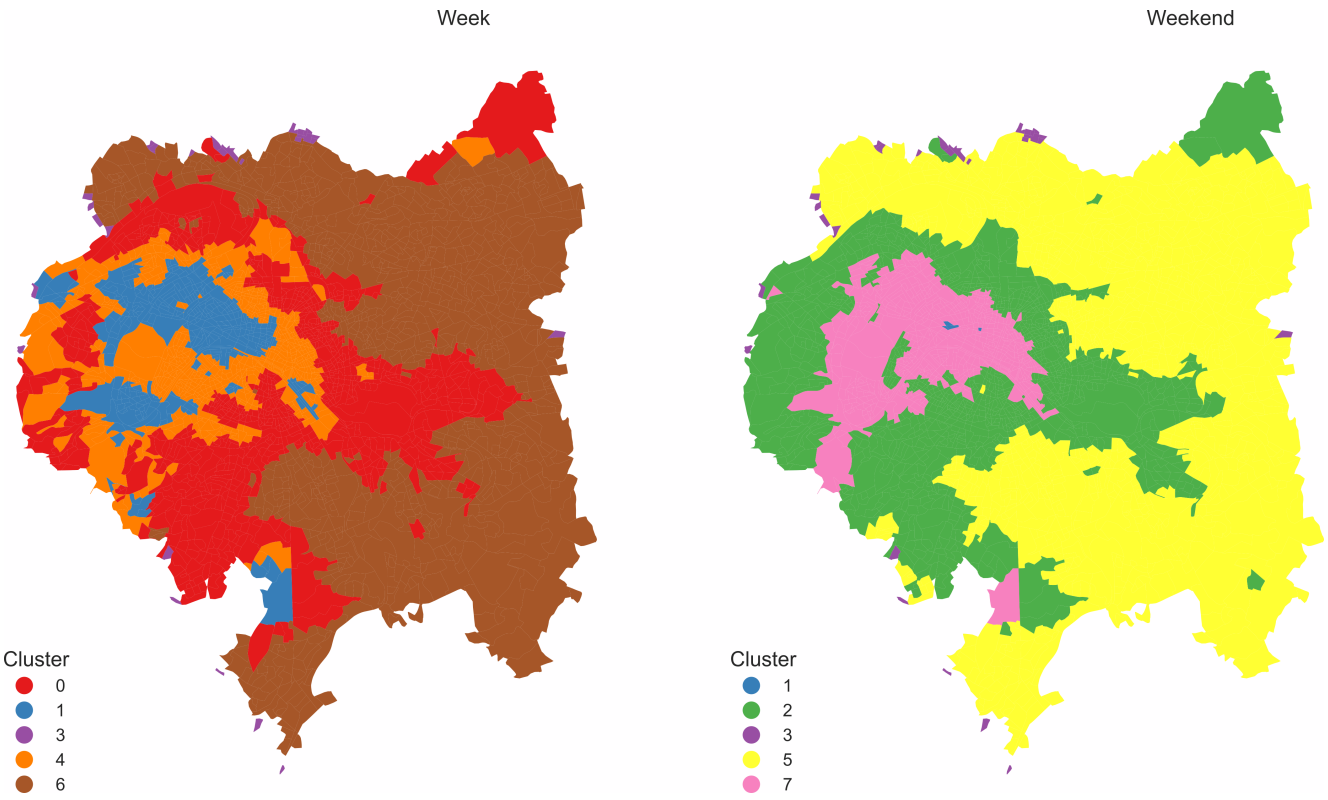

Figure 1: Cluster maps. Paris' clusters using the *IRIS2000* geometry [18]. Here, the maps of Paris are coloured by cluster for the (left) 'week' (Monday–Thursday) and (right) 'weekend' (Friday–Sunday). The clusters are given in the legend. These data are following preprocessing taken from the `NetMob23` data set [16]. We note here again, that, for Paris, cluster one and cluster three are shared but that cluster one on the weekend is only based on four cells; this is important when considering the histograms below.

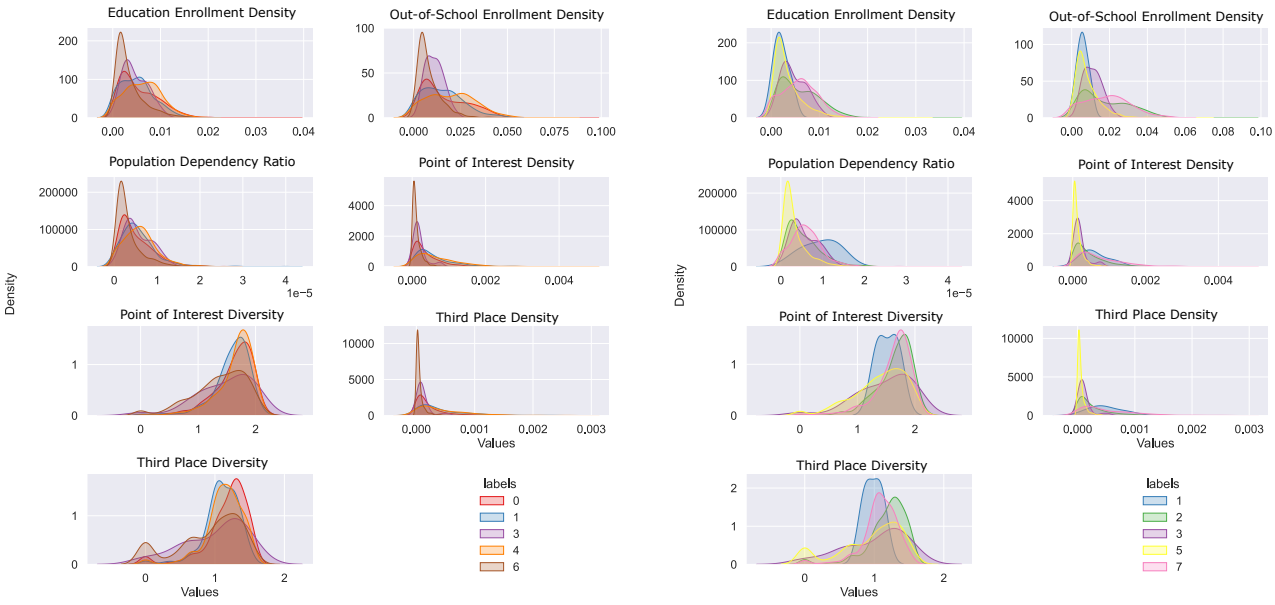

Figure 2: Cluster attribute histograms. Paris' clusters uysing the *IRIS2000* geometry [18]. Here, histograms show each independent variable that was included in a series of univariate analyses. For each variable, we partition the data giving each cluster. These data are following preprocessing taken from the `NetMob23` data set [16]. Note that, for Paris, cluster one and cluster three are shared but that cluster one on the weekend is only based on four cells.

Table 3: Moran's I: the minimum and maximum values from the clustering analysis for each city and each period in the analysis: 'week' and 'weekend'. We include here the Moran's I statistic, the std error of the value and the associated p-value. p-values are shown using their significance stars, where [0 - 0.001] is '***', [0.001 - 0.01] is '**', [0.01 - 0.05] is '*', [0.05 - 0.1] is '.', and [0.1 - 1.0] has no symbol. These data are following preprocessing taken from the `NetMob23` data set [16]. Note that the app-usage data often shows spatial clustering.

Temporal aggregation	City	Moran's I	STD	App feature	p-value
week	**Lyon**	0.48	18.12	Adult Content	***
	Marseille	0.11	3.67	Adult Content	***
	Paris	0.37	34.12	Adult Content	***
weekend	**Lyon**	0.51	19.25	Adult Content	***
	Marseille	0.08	2.76	Adult Content	***
	Paris	0.36	33.4	Adult Content	***
week	**Lyon**	0.84	31.77	Web Services	***
	Marseille	0.84	27.8	Web Services	***
	Paris	0.87	80.86	Web Services	***
weekend	**Lyon**	0.88	33.46	Web Services	***
	Marseille	0.82	27.22	Web Services	**
	Paris	0.88	81.53	Web Services	***

These values indicate that there is spatial clustering, however, we fully account for this in our main analyses. We tested whether spatial models were required for each of the models in the spatial modelling analysis. For this, we used the Lagrange multiplier calculations and the Akaike information criterion (AIC). We found that the Lagrange multiplier tests and the AIC tests pointed toward the use of SAR modelling. Because of this, we use this method going forward through the analysis.

Firstly, and more generally, we found some differences and some similarities between the week and weekend and across each city (see Figure 3). A considerable number of the significant results contain a work- or job-based use. Video conferencing, remote desktop, and professional development. It is however also important to highlight how a large number of variables are not statistically significant in the models, which we will explore further going forward.

Secondly, our aggregate model group is found to be quite similar to the Paris model group in both the significant variables, the strength of the effects, and the direction of the effects (see Figure 3). This is likely to be due to the influence of the Paris model group as it makes up the largest amount of data points. The aggregate model group shows quite similar results across the week and weekend with weak effects. In the aggregate model group again, the strongest effect was in the professional networking model where a negative effect was found for the education enrollment density variable ($Total$ = -2816.1062, p = 0.01, pr2 = 0.97). This suggests that higher education enrollment density is associated with lower levels of professional networking app usage. The relationship has a spatial component and is found to negatively influence surrounding cells. This could be because these areas have a strong focus on education and so lower engagement with professional networking apps. This was only found in the week. At the same time, there was a positive relationship between the dependency ratio and professional networking apps ($Total$ = 2783.73, p = 0.01, pr2 = 0.97). A higher ratio suggests a relatively larger dependent population (youth and the elderly were combined). As mentioned, these areas could be more family-orientated and the individuals within them could be using flexible work arrangements or parents are involved with professional development albeit not necessarily by educational means as this would conflict with the previous result. This was only found in the week which again points to caregiving families.

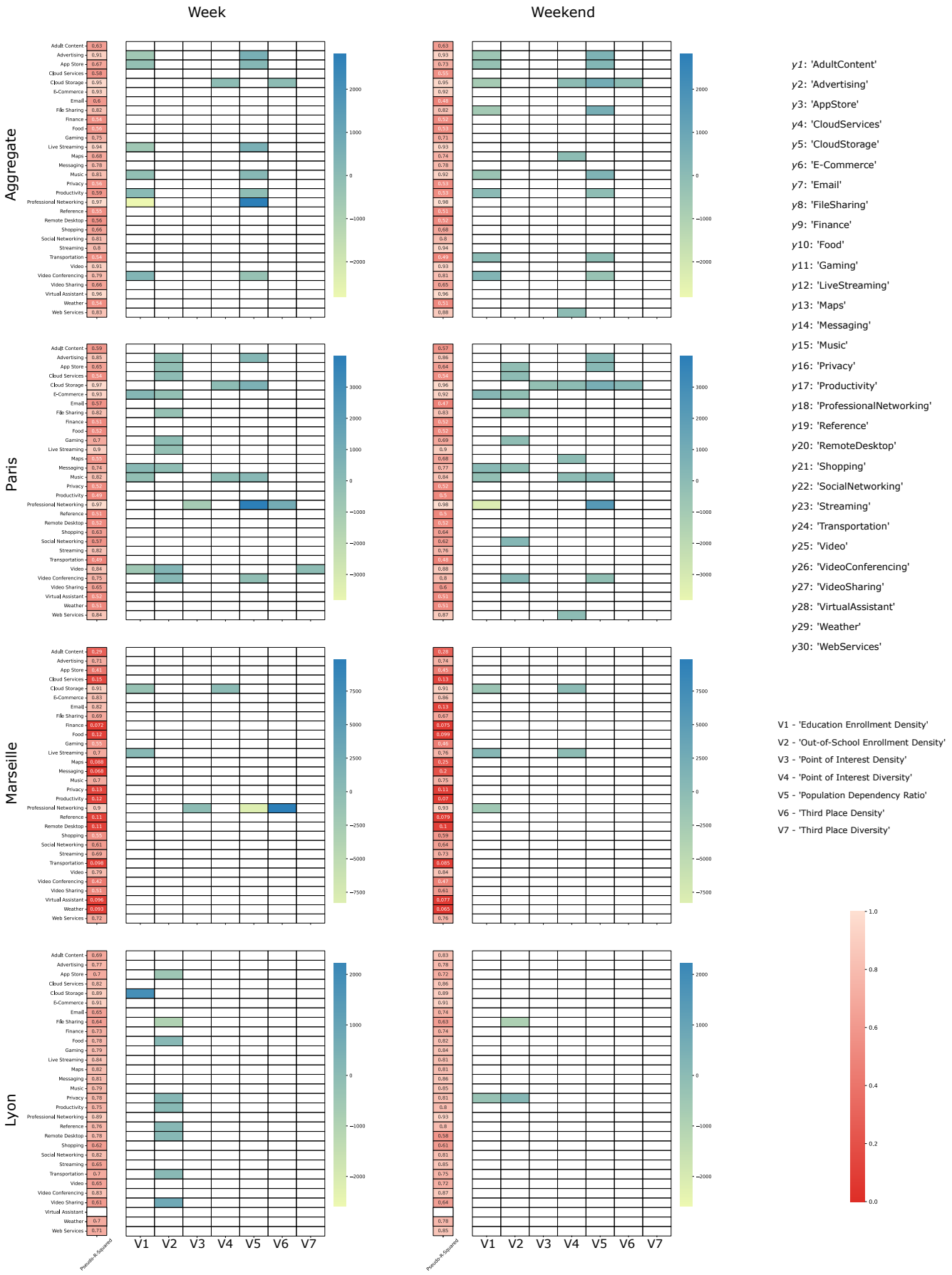

Figure 3: Univariate Model Heatmaps: Each facet represents a city (Aggregate, Paris, Marseille, Lyon) and period (Week, Weekend). The y-axis displays models alphabetically, while the x-axis shows analysis variables. Coefficients are color between yellow and blue according to the value, with white squares indicating non-significant results (i.e., where $p > 0.05$). Pseudo-R-squared values are presented in shades of red. Data preprocessing was applied to the `NetMob23` data set [16].

Going through each of the city model groups, in the Paris model group, and contrasting against the aggregate model group, there was no negative result for the professional networking model against education enrollment in the week, but there was on the weekend ($Total$ = -3181.34, p = 0.02, pr2 = 0.98; see Figure 3); this suggests perhaps that people are using leisure time to engage with professional development. Furthermore, for the professional networking model, the same positive result was found for the population dependency ratio in the week ($Total$ = 3989.16, p = 0.01, pr2 = 0.97) and this matched the weekend too ($Total$ = 2230.39, p = 0.04, pr2 = 0.97). This is perhaps further evidence that individuals use the weekend for professional development, but here individuals use the week and weekend equally. In the Marseille model group, the professional networking model had the strongest effect (see Figure 3). This time there was, instead of a positive effect, there was a negative effect in the week for the population dependency ratio variable ($Total$ = -8291.17, p = 0.01, pr2 = 0.9) and there was a positive relationship between professional networking and third-place density ($Total$ = 9851.27, p = 0.004, pr2 = 0.9). Both effects were found during the week, and neither was the case on the weekend. This means that, in the week, places with a higher dependency ratio, people may use professional networking apps less. The result for the weekend is not significant, which is interesting as it suggests that the dependency ratio does not appear to have a discernible impact on professional networking app usage. Regarding third places in the Marseille model group, this suggests that people may use third places for work-type networking activities. Again, this is only true of the week. In the Lyon model group, many models contained significant results for out-of-school enrollment density; however, these were only weak effects. The strongest effect was for the cloud storage app model: there was a positive effect between cloud storage and education enrollment ($Total$ = 1892.52, p = 0.04, pr2 = 0.89), this was for the week. The consistency of the professional networking model is not found in the Lyon group. The positive effect between cloud storage and education enrollment suggests people use cloud storage for education.

Lastly, across all model groups, we find that our pseudo-r-squared values are high despite having only a few significant results (see Figure 3). This is not necessarily unexpected due to the spatial regression models also capturing the spatial dependencies in the data and not fully explained by the independent variables.

5 Discussion

In this study, we used a large data set of app-usage data to study urban vibrancy in the three largest cities in France. We asked how urban features are related to urban vibrancy measured across an array of app-usage features. We added greater depth to the social aspect of the analysis by including socioeconomic indicators likely to be influential over urban vibrancy differences, and mediated by the social fabric of cities. We suggest that digital signatures, as measured by app usage, may exhibit interesting spatial properties, particularly in relation to the presence of specific urban features, such as third places, and for different demographic groups. We tested this via two computational approaches: (1) we use an HDBSCAN approach to maintain the structure of the apps together and to understand the cluster contents and most influential aspects, and (2) we use a series of univariate spatial models to test each of the variables separately whilst accounting for spatial autocorrelation and to understand how app usage interacts with neighbouring spatial units. Our main aim has been to reveal the link between urban vibrancy, urban features, and socioeconomics, so we used a collection of relevant variables for urban features known to be important in terms of urban vibrancy, and we integrated them with socioeconomic indicators.

We have uncovered some interesting findings. We have shown that it is possible to use highly detailed and high-frequency app usage data with socioeconomic indicators and geospatial data to observe potential spatial segregation in cities. We do this with high pseudo-r-squared values that indicate generally that our models have high predictive power and goodness-of-fit. We note

here again that pseudo-r-squared were high with only a few significant results. Going forward, we will investigate this further to better understand these results. Our HDBSCAN clustering analysis found distinct spatial clusters using the app usage data (Figure 1). These clusters seemed to have the most detail in the Paris cluster group. This could be because of the size of the data for Paris (see Table 2). We tested the attributes of the clusters compared to each of the variables (see Figure 2). We found some differences in the clusters, particularly when comparing the distribution of some urban features and demographic variables across the clusters. We carried out a number of spatial analyses (see Figure 3). Generally, we found that the results often indicated work-type behaviours but could also observe education and caregiving.

Furthermore, in the professional networking model, we observe a negative impact on the education enrollment density variable–as the density of education enrollments increases, there may be a decrease in the usage of professional networking apps; however, we identify a positive impact on the dependency ratio–as the density of dependency ratio increases, an increase in the usage of professional networking apps is observed. However, as the dependency ratio increases, increased usage of professional networking apps occurs indicating that individuals in these areas may use weekends for professional development and networking, possibly to improve their career prospects and support their dependents. There was a positive effect between cloud storage and education enrollment in Lyon. This may suggest people use cloud storage within their educational practices. Additional research is needed to draw definitive conclusions.

We also would like to highlight and discuss some of the limitations of the analysis. We used the data as a proxy for vibrancy and, despite the size of the data set, we should consider that the data do not encompass the entire population of each city in the analysis. We note here that our aggregate model seemed to be influenced heavily by Paris. Using more cities would help relieve this relationship. We used a low-resolution temporal aggregation that divided the data into week and weekend types. Though this enabled us to observe broad differences, it undoubtedly masked a great amount of detail. We suggest that using the time-series data within a time-series clustering approach might highlight more detail in the clusters. Our spatial aggregation used the *IRIS2000* cells as our main geographic reference. Using areal interpolation for the *IRIS2000* data could yield a larger sample size and give greater predictive performance.

In this study, we considered modelling urban vibrancy across multiple analyses including clustering approaches and a series of spatial regression models for each app. Our results add further evidence for the importance of using computational approaches for understanding urban environments, the use of sociological concepts in computational social science and for understanding urban vibrancy in cities.

References

[1] Michael Batty. “The Size, Scale, and Shape of Cities”. In: *Science* 319.5864 (Feb. 2008), pp. 769–771.

[2] Steven P. French, Camille Barchers, and Wenwen Zhang. “How Should Urban Planners Be Trained to Handle Big Data?” In: *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*. Ed. by Piyushimita (Vonu) Thakuriah, Nebiyou Tilahun, and Moira Zellner. Springer Geography. Cham: Springer International Publishing, 2017, pp. 209–217.

[3] Jan Nijman and Yehua Dennis Wei. “Urban Inequalities in the 21st Century Economy”. In: *Applied Geography* 117 (Apr. 2020), p. 102188.

[4] Yu Zheng et al. “Urban Computing: Concepts, Methodologies, and Applications”. In: *ACM Transactions on Intelligent Systems and Technology* 5.3 (2014).

[5] Jens Kandt and Michael Batty. "Smart Cities, Big Data and Urban Policy: Towards Urban Analytics for the Long Run". In: *Cities* 109 (Feb. 2021), p. 102992.

[6] M. Batty et al. "Smart Cities of the Future". In: *The European Physical Journal Special Topics* 214.1 (Nov. 2012), pp. 481–518.

[7] Raj Chetty et al. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States *". In: *The Quarterly Journal of Economics* 129.4 (Nov. 2014), pp. 1553–1623.

[8] Heather E. Price. "Large-Scale Datasets and Social Justice: Measuring Inequality in Opportunities to Learn". In: *Research Methods for Social Justice and Equity in Education*. Ed. by Kamden K. Strunk and Leslie Ann Locke. Cham: Springer International Publishing, 2019, pp. 203–215.

[9] Patrizia Sulis et al. "Using Mobility Data as Proxy for Measuring Urban Vitality". In: *Journal of Spatial Information Science* 16 (June 2018), pp. 137–162.

[10] Federico Botta and Mario Gutiérrez-Roig. "Modelling Urban Vibrancy with Mobile Phone and OpenStreetMap Data". In: *PLoS ONE* 16.6 June (2021), pp. 1–19.

[11] Pengyang Wang et al. "Measuring Urban Vibrancy of Residential Communities Using Big Crowdsourced Geotagged Data". In: *Frontiers in Big Data* 4 (2021).

[12] Yang Yue et al. "Measurements of POI-based Mixed Use and Their Relationships with Neighbourhood Vibrancy". In: *International Journal of Geographical Information Science* 31.4 (Apr. 2017), pp. 658–675.

[13] Jiayu Wu et al. "Urban Form Breeds Neighborhood Vibrancy: A Case Study Using a GPS-based Activity Survey in Suburban Beijing". In: *Cities* 74 (Apr. 2018), pp. 100–108.

[14] Thomas R. Collins et al. *Spatiotemporal Gender Differences in Urban Vibrancy*. Apr. 2023. arXiv: 2304.12840 [physics].

[15] Jane Jacobs. *The Death and Life of Great American Cities*. Random House; 1961.

[16] Orlando E. Martínez-Durive et al. *The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography*. May 2023. arXiv: 2305.06933 [cs].

[17] OSM. *OpenStreetMap Contributors*. https://www.openstreetmap.org/. 2017.

[18] The National Institute of Statistics and Economic Studies. *The National Institute of Statistics and Economic Studies*. https://www.insee.fr/en/metadonnees/definition/c1523.

[19] Geoff Boeing. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks". In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139.

[20] Alexis Comber and Wen Zeng. "Spatial Interpolation Using Areal Features: A Review of Methods and Opportunities Using New Forms of Data with Coded Illustrations". In: *Geography Compass* 13.10 (2019), pp. 1–23.

[21] Claudia Bergroth et al. "A 24-Hour Population Distribution Dataset Based on Mobile Phone Data from Helsinki Metropolitan Area, Finland". In: *Scientific Data* 9.1 (2022), pp. 1–19.

[22] Esteban Moro et al. "Mobility Patterns Are Associated with Experienced Income Segregation in Large US Cities". In: *Nature Communications* 12.1 (July 2021), p. 4633.

[23] Zhuangyuan Fan et al. "Diversity beyond Density: Experienced Social Mixing of Urban Streets". In: *PNAS nexus* 2.4 (Apr. 2023), pgad077.

[24] Ramon Oldenburg and Dennis Brissett. "The Third Place". In: *Qualitative Sociology* 5.4 (1982), pp. 265–284.

[25] Leo W. Jeffres et al. "The Impact of Third Places on Community Quality of Life". In: *Applied Research in Quality of Life* 4.4 (2009), pp. 333–345.

[26] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423.

[27] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 160–172.

[28] Peter J. Rousseeuw. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65.

[29] Michael Don Ward and Kristian Skrede Gleditsch. *Spatial Regression Models*. Second edition. Quantitative Applications in the Social Sciences. Thousand Oaks, California: SAGE Publications, 2019.

[30] James P. Lesage and Manfred M. Fischer. "Spatial Growth Regressions: Model Specification, Estimation and Interpretation". In: *Spatial Economic Analysis* 3.3 (Nov. 2008), pp. 275–304.

[31] Luc Anselin, Ibnu Syabri, and Youngihn Kho. "GeoDa : An Introduction to Spatial Data Analysis". In: *Geographical Analysis* 38.1 (2006), pp. 5–22.

[32] Luc Anselin and Sergio J. Rey. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC, Chicago, IL, 2014.

[33] K. H. Hamed. "The Distribution of Kendall's Tau for Testing the Significance of Cross-Correlation in Persistent Data". In: *Hydrological Sciences Journal* 56.5 (July 2011), pp. 841–853.

Supplemental material

Table SI 1: Correlation between up-link and down-link for each app in the analysis. We calculate the Kendall's tau (τ) rank correlation coefficient and the associated p-value. We have separated the values by the temporal categories of the 'week' and 'weekend' in line with the analysis. We use Kendall's tau because of the suitability for spatial data [33]. The table captures the strength and direction of the relationship of each correlation. p-values are shown using their significance stars, where [0 - 0.001] is '***', [0.001 - 0.01] is '**', [0.01 - 0.05] is '*', [0.05 - 0.1] is '.', and [0.1 - 1.0] has no symbol. Missing values are shown with a horizontal line. The apps are in alphabetical order.

	Paris				Marseille				Lyon			
	Week		Weekend		Week		Weekend		Week		Weekend	
	τ	P	τ	P	τ	P	τ	P	τ	P	τ	P
Amazon Web Services	0.86	***	0.85	***	0.83	***	0.84	***	0.84	***	0.85	***
Apple App Store	0.96	***	0.96	***	0.95	***	0.96	***	0.95	***	0.94	***
Apple Mail	0.81	***	0.78	***	0.81	***	0.76	***	0.77	***	0.81	***
Apple Music	0.96	***	0.96	***	0.95	***	0.93	***	0.97	***	0.97	***
Apple Siri	0.98	***	0.98	***	0.98	***	0.98	***	—	—	—	—
Apple Video	0.97	***	0.97	***	0.96	***	0.96	***	—	—	—	—
Apple Web Services	0.94	***	0.94	***	0.89	***	0.91	***	0.9	***	0.91	***
Apple iCloud	0.93	***	0.92	***	0.91	***	0.89	***	0.93	***	0.92	***
Apple iMessage	0.81	***	0.78	***	0.8	***	0.82	***	0.77	***	0.72	***
Apple iTunes	0.95	***	0.96	***	0.95	***	0.95	***	0.86	***	—	—
Clash of Clans	0.96	***	0.96	***	0.96	***	0.95	***	0.95	***	0.91	***
DailyMotion	0.95	***	0.94	***	0.95	***	0.94	***	0.96	***	0.95	***
Deezer	0.98	***	0.97	***	0.94	***	0.93	***	0.98	***	0.98	***
Dropbox	0.78	***	0.71	***	0.74	***	0.7	***	0.53	***	—	—
EA Games	0.75	***	0.74	***	0.74	***	0.75	***	—	—	—	—
Facebook Live	0.95	***	0.94	***	0.96	***	0.96	***	0.96	***	0.95	***
Facebook Messenger	0.91	***	0.91	***	0.87	***	0.86	***	—	—	—	—
Facebook	0.9	***	0.89	***	0.88	***	0.88	***	0.92	***	0.9	***
Fortnite	0.87	***	0.88	***	0.87	***	0.87	***	—	—	—	—
Google Docs	0.7	***	0.67	***	0.66	***	0.6	***	0.74	***	0.66	***
Google Drive	0.95	***	0.95	***	0.95	***	0.95	***	0.95	***	0.95	***
Google Mail	0.85	***	0.82	***	0.78	***	0.8	***	—	—	—	—
Google Maps	0.96	***	0.96	***	0.97	***	0.96	***	—	—	—	—
Google Meet	0.97	***	0.98	***	0.97	***	0.96	***	0.96	***	0.94	***
Google Play Store	0.94	***	0.92	***	0.94	***	0.93	***	0.96	***	0.95	***
Google Web Services	0.88	***	0.87	***	0.88	***	0.86	***	0.89	***	0.89	***
Instagram	0.98	***	0.98	***	0.98	***	0.98	***	—	—	—	—
LinkedIn	0.99	***	0.99	***	0.96	***	0.96	***	0.99	***	0.98	***
Microsoft Azure	0.67	***	0.64	***	0.66	***	0.7	***	0.69	***	0.69	***
Microsoft Mail	0.87	***	0.84	***	0.86	***	0.75	***	0.84	***	0.79	***
Microsoft Office	0.82	***	0.79	***	0.81	***	0.77	***	—	—	—	—
Microsoft Skydrive	0.65	***	0.59	***	0.63	***	0.61	***	—	—	0.44	***
Microsoft Store	0.95	***	0.94	***	0.94	***	0.94	***	0.95	***	0.95	***
Microsoft Web Services	0.87	***	0.84	***	0.83	***	0.84	***	0.87	***	0.86	***
Molotov	0.91	***	0.9	***	0.89	***	0.88	***	0.83	***	—	—
Netflix	0.96	***	0.95	***	0.95	***	0.94	***	0.94	***	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—

Table SI 2: Continued from above.

	Paris				Marseille				Lyon			
	Week		Weekend		Week		Weekend		Week		Weekend	
	τ	P	τ	P	τ	P	τ	P	τ	P	τ	P
—	—	—	—	—	—	—	—	—	—	—	—	—
Orange TV	0.94	***	0.94	***	0.89	***	0.88	***	0.95	***	0.94	***
Periscope	0.89	***	0.85	***	0.9	***	0.88	***	0.91	***	0.86	***
Pinterest	0.91	***	0.9	***	0.94	***	0.94	***	0.93	***	0.93	***
PlayStation	0.94	***	0.94	***	0.95	***	0.93	***	—	—	—	—
Pokemon GO	0.89	***	0.89	***	0.9	***	0.92	***	—	—	—	—
Skype	0.75	***	0.69	***	0.71	***	0.7	***	0.72	***	0.7	***
Snapchat	0.89	***	0.87	***	0.92	***	0.9	***	0.9	***	0.89	***
SoundCloud	0.95	***	0.95	***	0.92	***	0.92	***	0.93	***	0.93	***
Spotify	0.95	***	0.96	***	0.93	***	0.93	***	—	—	—	—
TeamViewer	0.64	***	0.58	***	0.55	***	0.68	***	—	—	—	—
Telegram	0.76	***	0.71	***	0.75	***	0.75	***	0.75	***	0.35	***
Tor	0.84	***	0.82	***	0.87	***	0.81	***	0.8	***	0.48	***
Twitch	0.92	***	0.91	***	0.92	***	0.91	***	—	—	—	—
Twitter	0.97	***	0.97	***	0.96	***	0.95	***	—	—	—	—
Uber	0.95	***	0.95	***	0.94	***	0.93	***	0.93	***	0.58	***
Waze	0.97	***	0.96	***	0.97	***	0.96	***	0.98	***	0.57	***
Web Ads	0.93	***	0.92	***	0.95	***	0.93	***	0.95	***	0.82	***
Web Adult	0.9	***	0.9	***	0.88	***	0.86	***	0.57	***	0.59	***
Web Clothes	0.98	***	0.98	***	0.96	***	0.95	***	0.94	***	0.91	***
Web Downloads	0.91	***	0.89	***	0.92	***	0.9	***	0.92	***	0.92	***
Web Finance	0.95	***	0.94	***	0.94	***	0.95	***	0.93	***	0.93	***
Web Food	0.71	***	0.68	***	0.7	***	0.75	***	0.7	***	0.73	***
Web Games	0.83	***	0.82	***	0.82	***	0.78	***	0.87	***	0.84	***
Web Streaming	0.91	***	0.9	***	0.92	***	0.91	***	0.94	***	0.94	***
Web Transportation	0.9	***	0.9	***	0.94	***	0.94	***	0.95	***	0.95	***
Web Weather	0.91	***	0.89	***	0.88	***	0.88	***	0.92	***	0.92	***
Web e-Commerce	0.96	***	0.96	***	0.97	***	0.96	***	0.98	***	0.97	***
WhatsApp	0.85	***	0.84	***	0.86	***	0.87	***	0.89	***	0.89	***
Wikipedia	0.96	***	0.96	***	0.91	***	0.91	***	0.86	***	0.88	***
Yahoo Mail	0.82	***	0.79	***	0.81	***	0.78	***	0.81	***	0.77	***
Yahoo	0.85	***	0.83	***	0.82	***	0.8	***	0.85	***	0.82	***
YouTube	0.93	***	0.92	***	0.91	***	0.91	***	0.95	***	0.94	***