



ARTICLE



<https://doi.org/10.1057/s41599-021-00822-w>

OPEN

Unraveling the association between socioeconomic diversity and consumer price index in a tourism country

Yan Leng ^{1✉}, Nakash Ali Babwany² & Alex Pentland³

Diversity has tremendous value in modern society. Economic theories suggest that cultural and ethnic diversity may contribute to economic development and prosperity. To date, however, the correspondence between diversity measures and the economic indicators, such as the Consumer Price Index, has not been quantified. This is primarily due to the difficulty in obtaining data on the micro behaviors and macroeconomic indicators. In this paper, we explore the relationship between diversity measures extracted from large-scale and high-resolution mobile phone data, and the CPIs in different sectors in a tourism country. Interestingly, we show that diversity measures associate strongly with the general and sectoral CPIs, using phone records in Andorra. Based on these strong predictive relationships, we construct daily, and spatial maps to monitor CPI measures at a high resolution to complement existing CPI measures from the statistical office. The case study on Andorra used in this study contributes to two growing literature: linking diversity with economic outcomes, and macro-economic monitoring with large-scale data. Future study is required to examine the relationship between the two measures in other countries.

¹ McCombs School of Business, The University of Texas at Austin, Austin, TX, USA. ² Independent Scholar, Toronto, ON, Canada. ³ Media Lab, MIT, Cambridge, MA, USA. ✉email: yan.leng@mcombs.utexas.edu

Introduction

Diversity is exceedingly valuable in modern society (Puritty et al., 2017). Empirical evidence relates diversity to tangible benefits, such as productivity (Sax, 2014; AlShebli et al., 2018) and innovation (Ottaviano and Peri, 2006) for organizations and nations (Galinsky et al., 2015). Diversity has attracted the interests from diverse fields. A commonly accepted view in cognitive science is that cognitive diversity enables the exchange of valuable information, thereby increasing creativity. Similarly, sociologists find that diverse ties provide greater access to social and economic opportunities (Eagle et al., 2010). There exist different opinions in economics; some claim that diversity measures predict economic growth (Montalvo and Reynal-Querol, 2005); some argue for a negative impact due to resource allocation between groups (Alesina and Ferrara, 2005); some argue that ethnic diversity deflates price bubbles (i.e., financial failures) (Levine et al., 2014); some argue that a diverse workforce (e.g., gender and racial diversity) is generally beneficial to corporate profits and earnings (Wright et al., 1995; Herring, 2009). Neighborhood ethnicity diversity has also been shown to have different effects on housing price (Macpherson and Sirmans, 2001). The enthusiasm from various disciplines highlights the importance of understanding diversity. To date, however, the correspondence between micro-level diversity and macro-level economic indicators (e.g., Consumer Price Index, measuring changes in the cost of purchasing a fixed basket of goods (Stigler, 1961)) has not been quantified. This gap is primarily due to the lack of data on both micro-behavior and economic indicators.

Macroeconomic indicators are essential for economists and policymakers to discern expansions and contractions in the near future (Bok et al., 2018). A couple of studies focus on predicting inflation and CPI using the data collected from the financial markets (Monteforte and Moretti, 2013; Modugno, 2013; Bañbura and Modugno, 2014). Establishing the relationship between micro-behaviors and macro-economic indicators is invaluable in monitoring the economic and social systems. Large-scale behavioral data can illuminate social phenomena and economic processes (Bok et al., 2018; Lazer and Radford, 2017). Traditional data collection methods are time-consuming, expensive to obtain, and vulnerable to sampling error (Bok et al., 2018). Hence, policymaking can be improved with high spatial-temporal resolution indicators comparing with statistics with substantial publication lags and limited contemporaneous information. Comparatively, large-scale behavioral data has high geographic and temporal granularity and is cost-effective (Lazer and Radford, 2017). Using massive data to approximate macro-indicators is especially promising in developing countries, where reliable data on economic livelihoods remains scarce (Blumenstock, 2016). There have been several successful attempts in the literature. To name a few, satellite and survey data have been combined to predict local economic outcomes in five African countries by applying a deep learning framework (Jean et al., 2016). Mobile phones have been combined with the environmental data to predict the Global Multidimensional Poverty Index (MPI) based on Gaussian Process regression (Pokhriyal and Jacques, 2017). The MIT Billion Prices Project used online prices to construct daily CPI in multiple countries Cavallo and Rigobon (2016).

Mobile phone data is especially promising in studying social and economic issues due to its high penetration rate and wide-availability (Leng et al., 2021, 2017). By January 2019, 5.1 billion users globally had mobile phones, with a penetration rate of 67%. Among them, 4.39 billion people had access to the internet (Kemp, 2019). More than 20 mobile phone companies have donated their proprietary information to developing big data solutions for social good (Kemp, 2019; Bakker et al., 2019). Epidemiologists have explored call records to combat diseases

(Maxmen, 2019), such as malaria in Kenya (Wesolowski et al., 2012), dengue in Pakistan (Wesolowski et al., 2015), and cholera in Haiti (Rinaldo et al., 2012). Researchers have also used mobile phone data to extract macro-level indicators for touristic events performances (Leng et al., 2016a). During the COVID-19 pandemic, we have seen an increasing number of applications using mobile phones to do contact tracing (Oliver et al., 2020). Therefore, mobile phone data has a nature appeal to unveil the relationships between micro-level behaviors (e.g., diversity) and economic statistics (Leng et al., 2016a). In this paper, we present a case study to explore the relationship between the travel demand based diversity measures extracted from mobile phone data and one of the most important macroeconomic indicators in this tourism country, namely the Consumer Price Index.

To the best of our knowledge, our study constitutes the first attempt to study the association between tourists-based diversity and the CPI in a tourism country. We couple the most complete national communication data with the CPI statistics collected by the Andorran government. Although the nature of the data limits the ability to establish a causal relationship, we can explore the association between the micro-level diversity and the region economic indicator, both tourism and non-tourism related. Specifically, we investigate the relationship between socio-demographic diversity extracted from mobile phone data surrounding different types of POIs¹ and the CPI of different industries. Our results demonstrate strong predictive relationships between diversity measures of nationalities and income and the Consumer Price Index of different sectors. We build statistical models to predict the CPIs of different categories in a European tourism country, Andorra. Finally, we estimate the CPI on a daily basis and at the cell tower levels, improving the temporal availability and spatial precision of CPI estimations. This associate shows that diversity measures on tourists can be used to predict the macroeconomics in a tourism country.

This paper proceeds as follows. Section “Methods” covers the setting and data of this study. Section “Results” presents the results on the relationships between CPI and the diversity measures from mobile phone data. Section “Nowcasting and mapping” CPI presents daily nowcasting and spatial maps of the CPI measures. Section “Discussion” summarizes and concludes the paper.

Methods

Call detail records. We use passively collected behavioral data, call detail records (CDRs), in a European country, Andorra. The economy of Andorra heavily relies on tourism. The population of Andorra is only 85,000, while it attracts 10.2 million international visitors annually (Cia.gov., 2012). The high volumes of tourists makes Andorra an especially interesting country to study socio-demographic diversity.

Mobile carriers initially collect CDRs for billing purposes; hence, this data widely exists in almost every country in the world. The spatial-temporal resolution of this data is high compared to traditional surveys and has the most substantial penetration rate among all passively collected data. The data is recorded when users make phone calls, send short message services (SMS), and use internet data services. It contains information on the longitude and latitude of the cell tower, the timestamp of the transaction, the registry country of the SIM card, and other characteristics about the phone (e.g., the brand, the vendor, the model, and the system of the phone). The CDRs were collected from July 2014 to August 2016 the only mobile carrier in Andorra. Hence, the coverage of mobile phone data is 100% in our study, meaning that we have all the mobile phone data of individuals who visited Andorra.

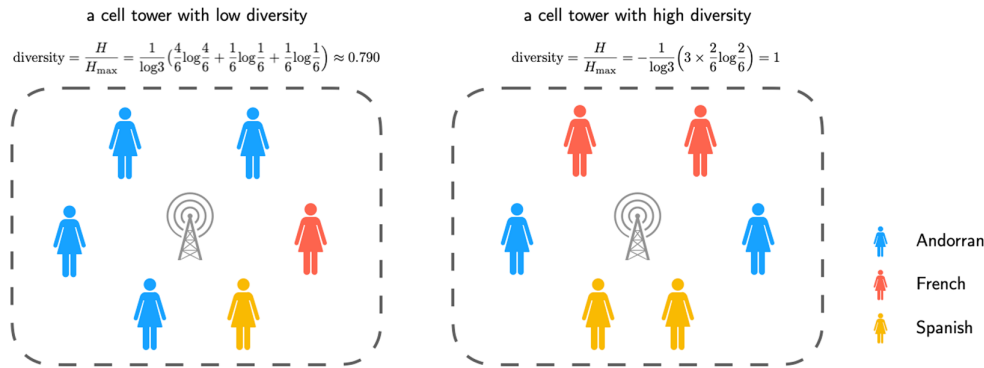


Fig. 1 Diversity at cell tower. We illustrate the diversity measure on nationality. Assume individuals connecting to the cell tower come from one of the countries, {Andorra, Spanish, French}. The left panel—dominated by Andorrans—is less diversified than the right panel.

Categories of cell towers. There are one hundred cell towers in Andorra, each covering an area of 250 m to 2 km in radius. We use the Voronoi tessellation to approximate the mobile tower coverage (Fortune, 1987). We manually label the cell towers to eight categories of Points of Interests (POIs), including wellness, leisure, shop, gastronomic, nature, event, culture, and others. Each tower may be associated with more than one POI category.

Diversity measures. The diversity measure is defined using two types of information, namely, nationality and approximation of disposable income. We compute the diversity measures at each cell tower and aggregate according to the eight kinds of POIs. We quantify diversity of cell tower as a function of Shannon entropy. According to Stirling (2007), there are three types of diversity measures, namely balance, variety, and disparity. The Shannon entropy we employ in the paper captures balance. Balance measures the pattern of apportionment of tourists from different origin countries. Specifically, Shannon entropy in our case captures the evenness or concentration of tourists origins. It measures how likely a tourist from a certain origin interacts with tourists from another country. We believe the types of interactions, how likely a foreigner interacts with tourists from another country and how likely a local Andorran interacts with tourists from different countries, contribute to higher CPI. Variety and disparity are not appropriate in our study for the following reasons. Variety is the number of categories into which system elements are apportioned. However, a simple enumeration of the number of countries tourists travel from cannot capture whether there are just one tourist from one country. Assume that there are one tourist from each country around the world in Andorra, varieties are large, however, cannot capture diversity. The other type is disparity, which refers to how each tourist origin can be distinguished from another. It measures how different are the foreign tourists from another country in Andorra. This is not appropriate in our setting in capturing diversity, as our goal is not to differentiate tourists from one country of origin from another.

Next, we describe how we compute the diversity measure in this study,

$$H = \sum_{j=1}^J p_j \log(p_j), \quad (1)$$

where p_j is the proportion of individuals who belong to category j , and J is the total number of categories. The diversity is computed at a daily level τ , where $\tau \in T$ and T is the set of days over the total observational period. The diversity of nationalities at cell tower i

on day τ , $D_{\text{nat},i,\tau}$ is formulated as,

$$D_{\text{nat},i,\tau} = \frac{1}{\log(K_{\tau}^+)} - \sum_{k=1}^K \frac{t_{i,k,\tau}}{T_{i,\tau}} \times \log \frac{t_{i,k,\tau}}{T_{i,\tau}}, \quad (2)$$

where $T_{i,\tau}$ is the total number of individuals connected to cell i during τ , $t_{i,k,\tau}$ is the number of individuals belonging to nation k who connected to cell tower i during τ . K_{τ}^+ is the number of distinct nationalities that appear at the cell tower on day τ . K is the total number of nations and $K = 10$ in our study. The nations consist of Andorra and other countries with frequent tourists to Andorra, including Spain, France, Netherlands, Belgium, Russia, the UK, Germany, Portugal, and others.

We use phone prices to approximate the disposable income of mobile phone users. We discretize phone prices into 14 categories, including [0, 20], [20, 30], [40, 50], [50, 100], [100, 150], [150, 200], [250, 300], [300, 400], [400, 500], [500, 600], [600, 700], [700, 800], [800, 900], higher than 900. All units are in USD. S is the number of phone price categories, and $S = 14$ in our case. The diversity of income at cell tower i on day τ , $D_{\text{inc},i,\tau}$ is constructed similarly using Shannon entropy and is normalized by the number of categories. Formally, it is calculated as:

$$D_{\text{inc},i,\tau} = \frac{1}{\log(S_{\tau}^+)} - \sum_{s=1}^S \frac{p_{i,s,\tau}}{T_{i,\tau}} \times \log \frac{p_{i,s,\tau}}{T_{i,\tau}}, \quad (3)$$

where S_{τ}^+ is the number of phone price categories with at least one users on day τ . $p_{i,s,\tau}$ is the number of individuals connected to cell tower i who belong to phone price category s on day τ .

A higher diversity measure implies that the cell tower attracts a more diversified (i.e., less uniform) population; see Fig. 1 for an illustration. The left figure is dominated by Andorrans (with a lower diversity measure), while the right plot is more diversified (with a higher diversity measure). We perform z-normalization on all of the predictors used in this study to allow for an easier comparison of variable importance.

Let $\mathcal{C}(b)$ be the set of cell towers with POI type b . $|\mathcal{C}(b)|$ is the number of cell towers with POI b . The average diversity of nationality at POI category b on day τ is:

$$\text{Div}_{\text{nat},b,\tau} = \frac{1}{|\mathcal{C}(b)|} \sum_{j \in \mathcal{C}(b)} D_{\text{nat},j,\tau}. \quad (4)$$

Similarly, the average diversity of income at POI category b is:

$$\text{Div}_{\text{inc},b,\tau} = \frac{1}{|\mathcal{C}(b)|} \sum_{j \in \mathcal{C}(b)} D_{\text{inc},j,\tau}. \quad (5)$$

Note that since one cell tower may be assigned to more than one POI category, they may contribute to more than one POI categories.

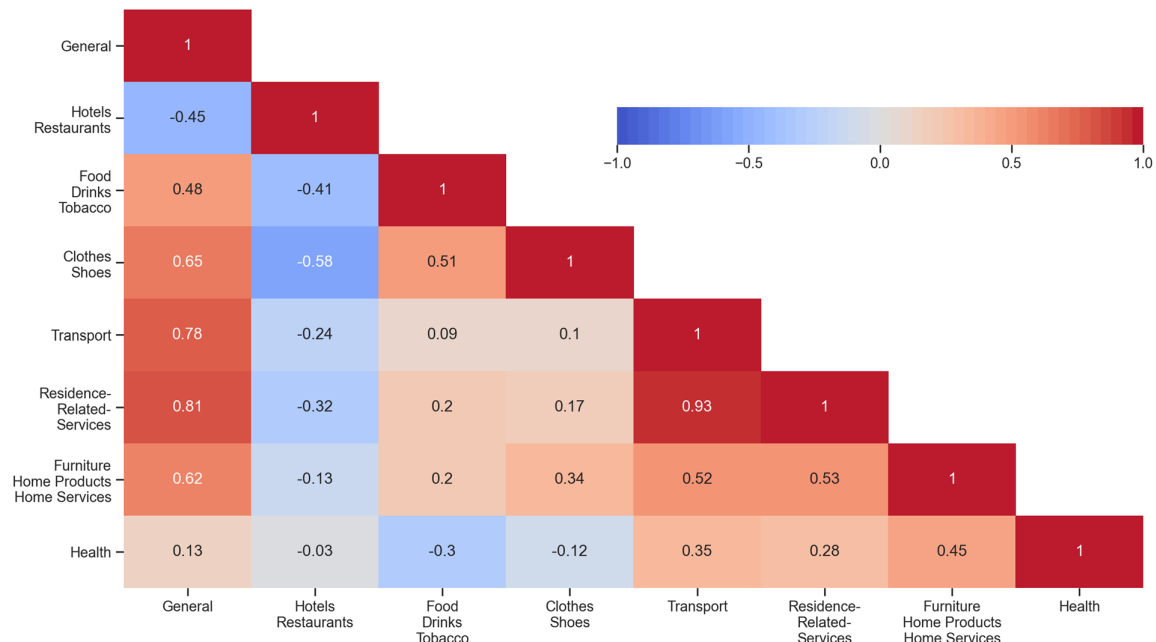


Fig. 2 Correlations between the CPIs of different industries. The darker the color, the stronger the correlation.

Consumer price index. We collect the monthly CPI from the Andorra Government Statistics website². The CPI measures we collected are a relative value compared with year 2001³. Other than general CPI, we also collect CPI measures in different industries. We segment them into tourism-related and residents-related CPIs:

- Tourism-related CPI: (1) hotels, cafes, and restaurants; (2) food, drink, and tobacco.
- Resident-related CPI: (1) transport; (2) clothes and shoes. (3) residence related services⁴; (4) public and social security administration; (5) furniture, products and services for home⁵; (6) health.

We focus on the change in CPI relative to the past month. The relative change in CPI in month $t+1$ relative to month t is defined as,

$$\Delta \text{CPI}_t = \frac{\text{CPI}_{t+1} - \text{CPI}_t}{\text{CPI}_t}. \quad (6)$$

In Fig. 2, we present the correlations between each pair of industries. Many of the resident-related CPIs (i.e., clothes, resident services, home products, and transport) are highly positively correlated with the general CPI. Among these industrial CPIs, the correlation between the CPI of transport and resident services is the highest among all pairwise correlations (correlation = 0.93, $p\text{-val} < 0.001$). Two of the tourism-related CPIs (i.e., food, drink and tobacco, and leisure and culture) correlate with general CPI to a lesser extent. Interestingly, the direction of the correlations is the opposite between the two. CPI of the hotel and restaurant industries negatively correlates with the general CPI, while the correlation between CPI of food, drink and tobacco, and general CPI is positive.

Results

Correlation between diversity and CPI. We first investigate the association between diversity and consumer price index (CPI), as illustrated in Figs. 3 and 4. We present the Pearson correlation in Fig. 3. The association between the diversity of income at leisure

and nature places is exceptionally high, with correlations of 0.805 ($p\text{-val} < 0.001$) and 0.775 ($p\text{-val} < 0.001$). We further illustrate the relationships via a scatter plot in the leftmost plot of the first row in Fig. 4. This pattern shows that we can easily predict general CPI using a single diversity measure mentioned above.

We observe that more diversity of nationality in the country predicts a deflation of the CPIs of (1) general ($r = -0.650$, $p\text{-val} < 0.001$), (2) food, drinks and tobacco ($r = -0.635$, $p\text{-val} < 0.001$), and (3) clothes and shoes ($r = -0.765$, $p\text{-val} < 0.001$), while it predicts inflation of the CPIs of hotels and restaurants ($r = 0.751$, $p\text{-val} < 0.001$). More diversity in income in the country predicts an inflation in (1) general ($r = 0.743$, $p\text{-val} < 0.001$), (2) transport ($r = 0.682$, $p\text{-val} < 0.001$), and (3) residence related services ($r = 0.650$, $p\text{-val} < 0.001$).

Besides, many of the diversity measures are highly predictive of CPIs in hotels and restaurants. For example, the diversity of nationality of the country and at shops associate with this diversity measure with strong correlation ($r = 0.75$, $p\text{-val} < 0.001$). We present the scatter plot in the two rightmost figures of the first row in Fig. 4. Additionally, four of the diversity of nationality measures (i.e., diversity of nationality at the country level ($r = -0.765$, $p\text{-val} < 0.001$) and shopping ($r = -0.750$, $p\text{-val} < 0.001$), food ($r = -0.785$, $p\text{-val} < 0.001$), and cultural POIs ($r = -0.756$, $p\text{-val} < 0.001$)) negatively associate with the CPI of clothes and shoes.

Diversity of income and nationality present different roles in predicting inflation and deflation, as can be seen from the contrary patterns. Diversity of income at different POIs positively correlates with general CPI. In contrast, the diversity of nationality presents a negative correlation. A similar contrary pattern shows up in most other CPIs: diversity of nationality negatively correlates with CPI of residence related services ($r = -0.401$, $p\text{-val} < 0.05$), and furniture, home products, and home services ($r = -0.392$, $p\text{-val} < 0.05$). In contrast, the diversity of income positively correlates with these CPI measures: diversity of nationality negatively correlates with CPI of transport ($r = 0.682$, $p\text{-val} < 0.001$), residence related services ($r = 0.650$, $p\text{-val} < 0.001$), and furniture, home products, and home services ($r = 0.395$, $p\text{-val} < 0.05$).

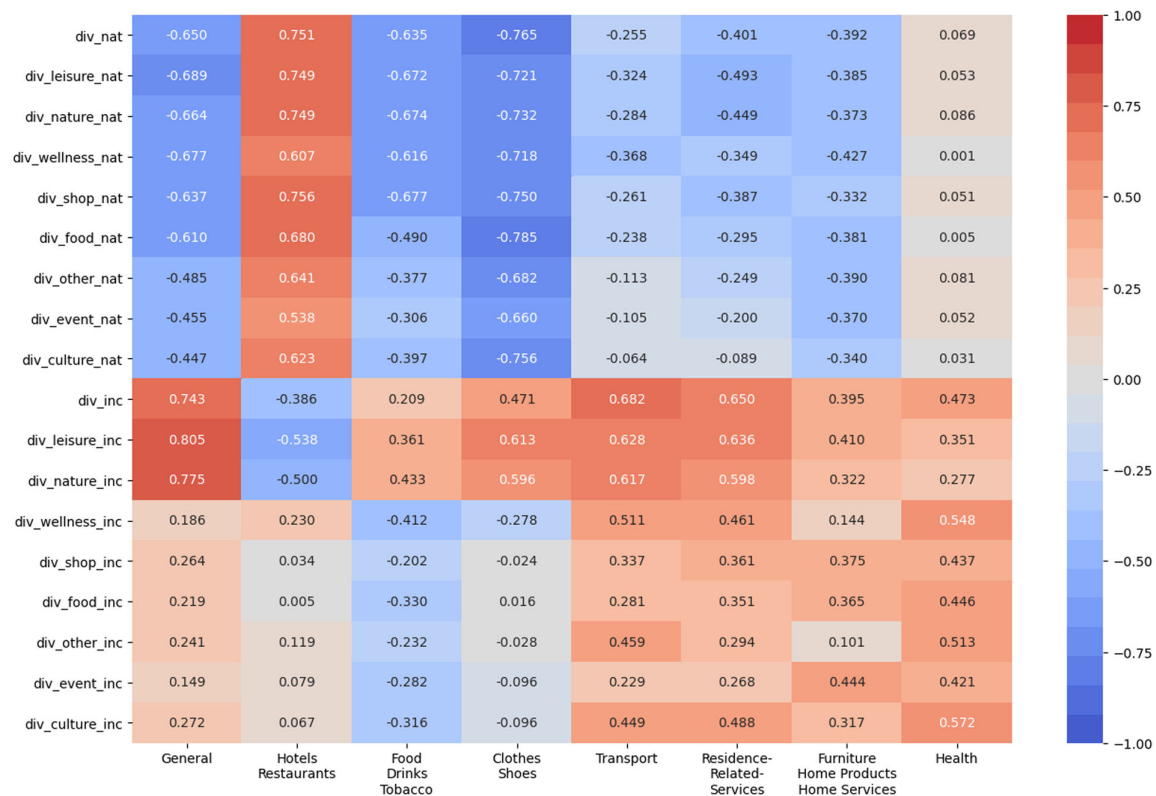


Fig. 3 Correlation between diversity measures and CPI. Heatmap of Pearson correlation in diversity measure (x-axis) and CPI (y-axis). The top nine rows are nationality-related diversity measures. The bottom nine rows are income-related diversity measures. The value corresponds to Pearson correlation with the legend shown on the right. Blue to red corresponds to correlation coefficients ranging from -1 to 1 .

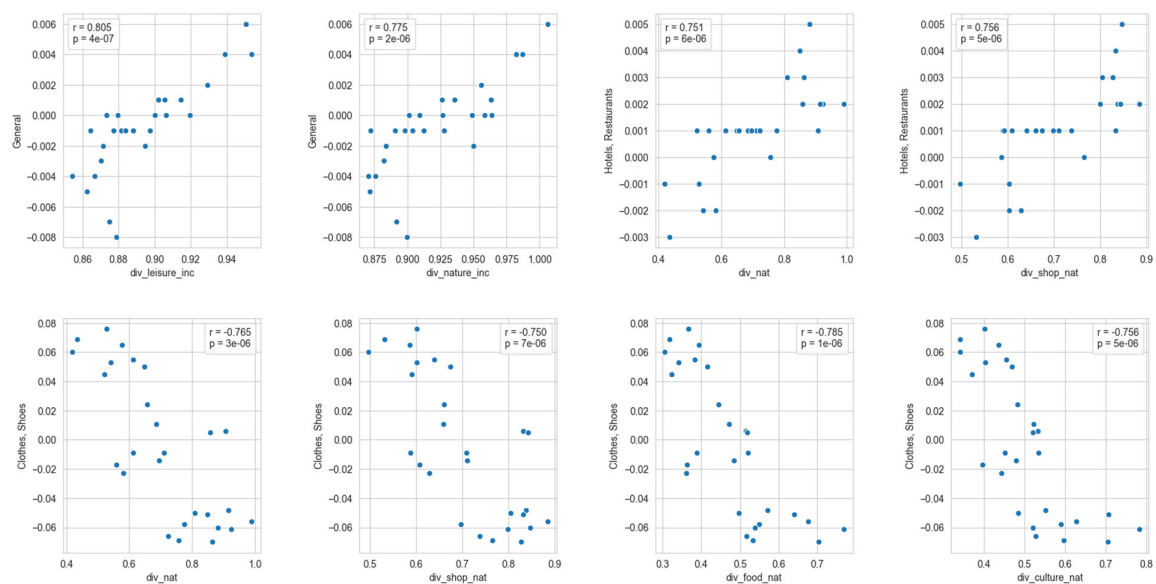


Fig. 4 Scatter plot of diversity measures and CPI measures of the top eight by the coefficients. The x-axis and y-axis correspond to diversity measures and CPIs, respectively.

A selective set of diversity measures. Next, we use the elastic net regression method to select the most important covariates for predicting each CPI measure, as shown in Figs. 5 and 6. This analysis helps us understand whether we can use a small number of diversity measures to reach reasonable predictive performances

on CPIs. We see that diversity of income in the country, and at leisure and nature POIs and the diversity of nationality at wellness POIs are more predictive of general CPI than other diversity measures. The three income-related diversity measures exhibit a positive relationship with general CPI, while the nationality

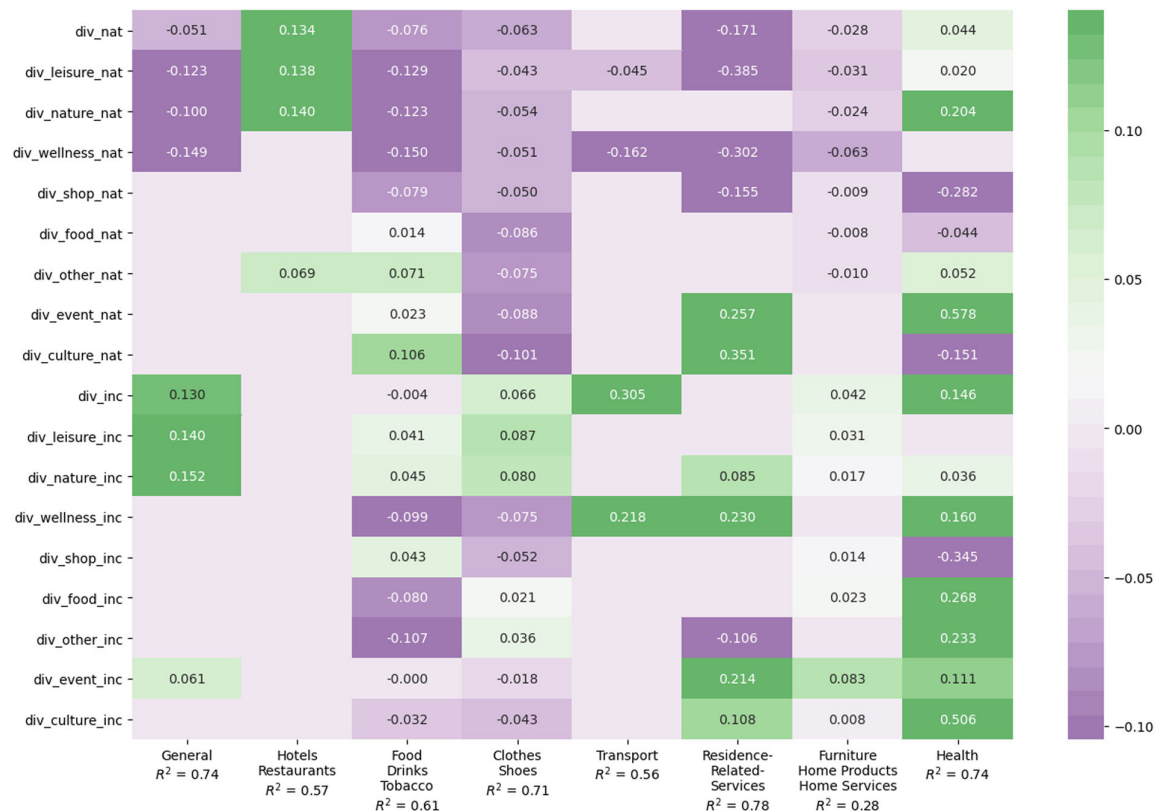


Fig. 5 Covariate selected by elastic net regression. We present the coefficients selected via elastic net regression. The value in the cell correspond to the coefficients in the selected model. The covariates with non-zero coefficients are presented.

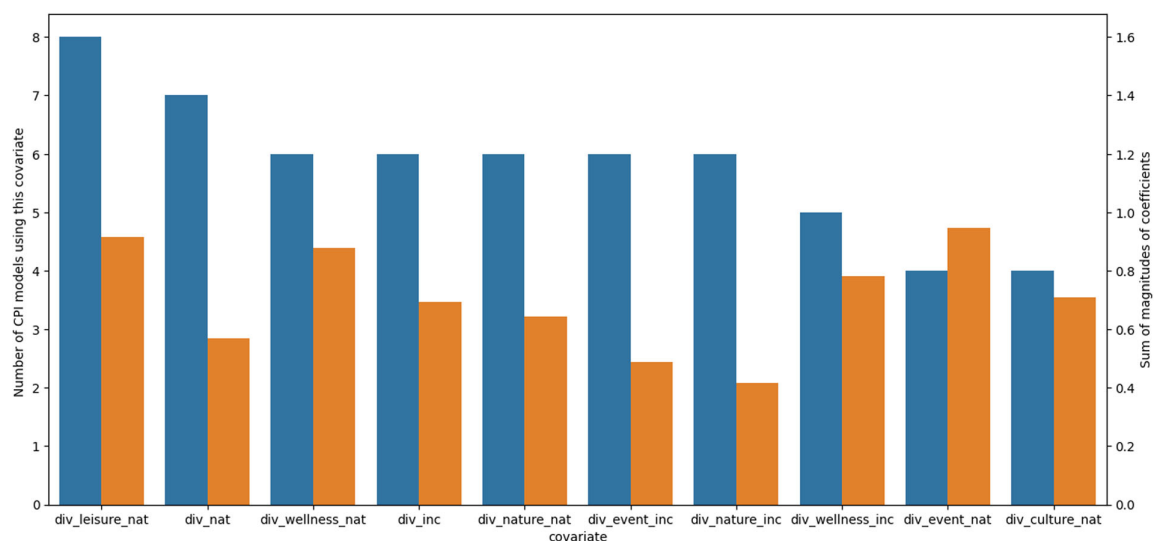


Fig. 6 Covariate importance. The number of models selected this covariate (the y-axis on the left) and the sum of coefficients in all models (the y-axis on the right). The x-axis corresponds to the top covariates selected by both axes.

related diversity measures exhibit a negative correlation. In terms of the CPI of hotels and restaurants, four covariates related to the diversity of nationality, including diversity of nationality in the country and at leisure, nature, and other POIs, are sufficient to achieve reasonable outcomes. All of the diversity measures contribute positively, similar to the correlation pattern shown in Fig. 3. For general CPI, eight diversity measures are selected to achieve

$R^2 = 0.74$, among which both diversity measures in the whole country and that at leisure and nature places are selected. Also, four diversity measures of nationality contribute negatively, and four diversity measures of income contribute positively to general CPI. This pattern—income diversity contributes positively, and nationality diversity contributes negatively—also shows up in models for transport (with $R^2 = 0.78$).

Four covariates are sufficient to predict the CPI of (1) transport; and (2) hotels and restaurants. Comparatively, all covariates provide additional predictive powers to nowcast the CPI of (1) clothes and shoes, and (2) food, drinks, and tobacco.

We summarize the importance of covariates in Fig. 5. We present the top ten important covariates by the number of models that contribute to and the sum of coefficients in magnitude. The diversity of nationality at leisure POIs and in the whole country are predictive for most CPIs. Diversity of nationality at the event, wellness, and leisure POIs are the most significant by the magnitude of coefficients in all POIs. Diversity measures at the shop, food, and other POIs are not as predictive to CPI measures comparing to the top ten diversity measures. This result indicates that diversities at leisure, nature, wellness, event, and culture POIs are more predictive for the CPIs in this tourism country. As data on more countries become available, future research can explore whether this pattern exists in other tourism countries.

Nowcasting and mapping CPI

Since the population's diversity can be computed at a much higher spatial and temporal resolution, we can nowcast⁶ the CPI with much finer granularity. Using the covariates selected via elastic net regression⁷, we nowcast Δ CPI at a daily level and at the cell tower level. This analysis further highlights the benefit of using mobile phone data: it allows the government to adjust economic policies timely, leading to a data-driven smarter city.

Temporal nowcasting on a daily basis. Regular updates on macroeconomic indicators are necessary to enable the federal government to (1) adjust historical data, to (2) escalate federal payments and tax brackets, and to (3) adjust rents and wages (Bok et al., 2018). To this end, our result helps to provide frequent data for policymakers to complement the infrequent macroeconomic statistics. This analysis demonstrates the potential of enhancing the timely proxy for CPI measures at the country and sectoral levels with mobile phone data. The nowcasting for daily CPI measures is shown in Fig. 7. We see that our model can capture the variations in the CPI measures. Most CPI indicators demonstrate a periodic pattern except for home product services, which is relatively flat, and health services showed an increasing trend in 2015. Our method captures the periodic pattern well, especially for clothes and shoes.

Mapping regional CPI. CPI is usually reported on a national scale, while our method can estimate such measures at a regional level. This regional nowcasting provides more insights for policymakers to design regional policies. In Fig. 8, we show the spatial distributions of the predicted general CPIs in June 2015 and January 2016. The regional variation shown in both plots demonstrates the regional differences in CPIs, which implies that the country-level CPI is not sufficient to capture the regional variations. We perform community detection on the cell towers according to the predicted CPI. We observe that even if spatially proximate cell towers belong to similar groups, we still see some farther-away cell towers being grouped. The regional nowcasting might be helpful for policymakers to design corresponding interventions to deal with the regional variations in CPI.

Discussion

Our paper reveals strong associations between diversity measures and CPI measures in an European country of Andorra, contributing to the growing literature of diversity in various disciplines. We find that the diversity of income at leisure and nature POIs alone is highly predictive of general CPI. Moreover, diversities of nationality at the country level and shops are highly

correlated with the CPI of hotels and restaurants. Interestingly, these two diversity measures negatively correlate with the CPI of clothes and shoes. Also, we use a statistical model to select a smaller number of covariates for predictions. Our result shows that socio-demographic diversities of tourists in Andorra and at some tourism-related (e.g., leisure, nature, and wellness) POIs are highly predictive of multiple sectoral CPIs. This result is useful when we cannot compute the diversity measures for many POIs when data is limited or POIs are sparse. Although the data cannot be used to show causality, the association suggests that diversity can be a strong predictor for CPI. Our finding provides empirical evidence to support the relationships between social structures (i.e., the diversity of individuals in a small region) and the CPI of different industries (both tourism and non-tourism related). Diversity of tourists may indicate that the service and tourism industry is attractive to tourists in different countries and with different income levels. This suggests the wellness of the economy. In addition, more international tourists may inflate the prices. All of these leads to changes in CPI.

Undoubtedly, new data and information technology can improve timely statistics in economics and monitoring society (Lazer and Radford, 2017). Our work represents an attempt to build predictive maps and daily predictions of CPI using the mobile phone data in a tourism-based European country, contributing to the burgeoning literature using big data to produce timely macroeconomic indicators. The strong association may be due to the strong relationships between behavior of tourists and the CPIs of this tourism country. The universal coverage of cell towers and the wide availability of CDRs makes it possible to predict CPI at high spatial and temporal resolutions in other countries. As behavioral data becomes more available, the high-resolution predicted economic indicators from this data could complement the static and lagged government statistics to help policymakers and economists make more informed decisions. Using travel-demand-based CDRs to provide CPI estimates of different industries to deliver accurate, high-resolution CPI maps offer a way to complement traditional statistical methods and provide regular updates in high spatial resolution in this tourism country. This study offers a framework to utilize human behavioral data by aggregating information at cell tower levels without revealing sensitive user information.

Our study is not without limitations and hence points out several future directions. We provide empirical evidence demonstrating the relationship between diversity and inflation (deflation) in a tourism country of Andorra. Andorra is an interesting case study as a tourism country, as it is highly internationally and attracts tourists from all around the world. This makes Andorra an especially interesting case study for diversity. We expect that such study can possible extend to other tourism countries or cities that attract international travelers. We use Andorra as a case study and this opens up opportunities for such analysis in other countries using mobile phone data. We leave the analysis to other contexts (non-tourism countries) for future work. As more data becomes available, we expect the same framework to be applied in different countries to examine the external validity of this study to more countries (tourism and non-tourism countries). The economic literature has laid out several micro-foundations to explain the forces underlying ethnic diversity and economic development Alesina and Ferrara (2005), related to individual preferences Alesina et al. (2000), individual strategies Alesina and Ferrara (2005), and production function (e.g., heterogeneity vs. innovation and productivity) Ottaviano and Peri (2006). More theoretical groundings may grow out of this work to explain the relationships between diversity and CPI, especially in a tourism country.

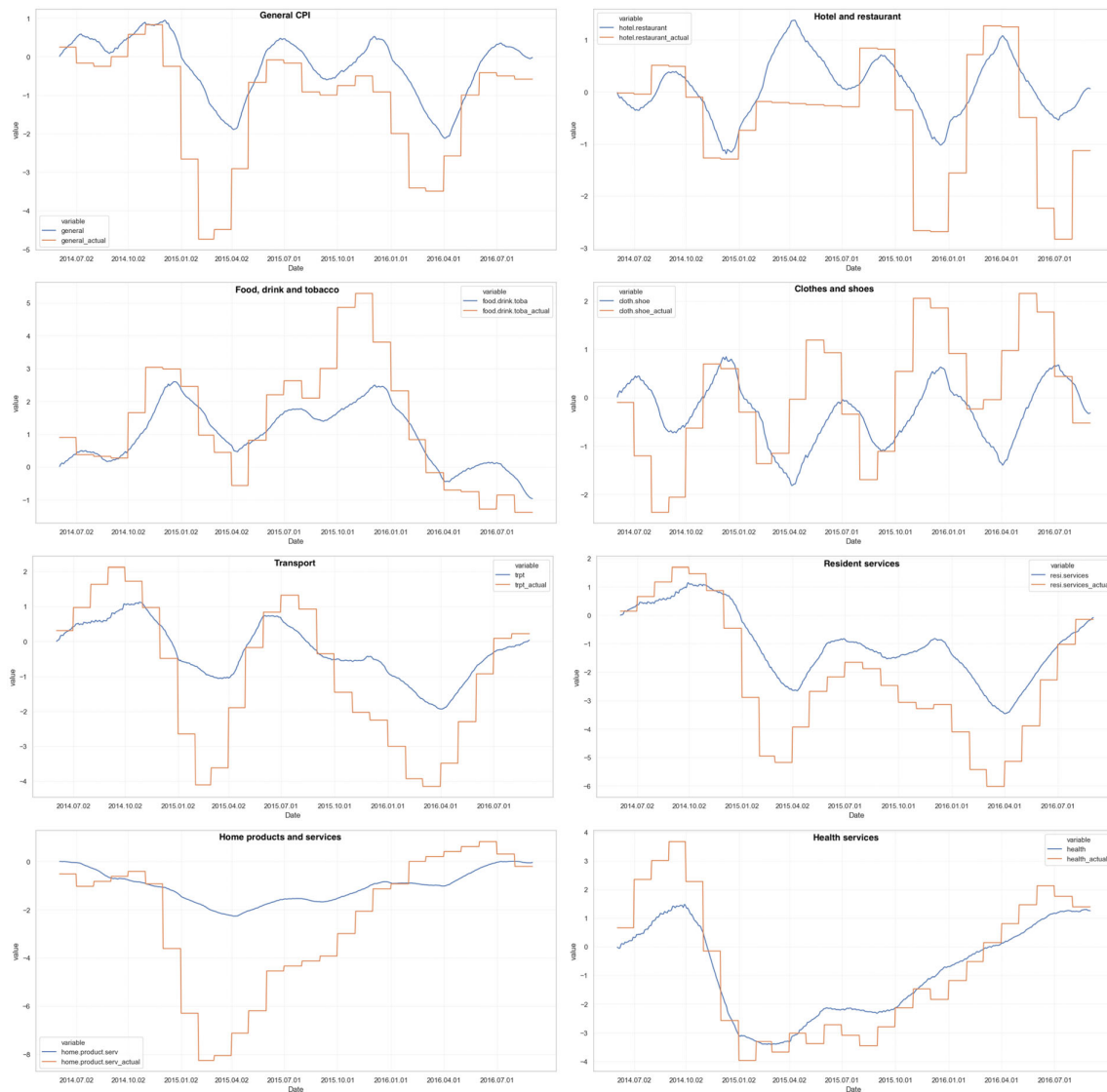


Fig. 7 Daily CPI nowcasting. The x-axis corresponds to the days. The y-axis is the predicted CPI (in blue) and the actual CPI (in orange), relative to the start of the period (June 1, 2014).

Uncovering the underlying mechanism connecting diversity and CPIs is not within the scope of this study. However, we provide some potential explanations. Future studies can unveil the causal mechanism. First, the diversity of tourists may indicate that the tourism industry of Andorra is attractive to diverse tourists (in different countries and with different income levels). The attractiveness to diverse cultures and income may predict a higher CPI. This also suggests the economic wellness of the country, which may, in turn, attract more tourists. Second, with many tourists from different nations, the Andorra tourism department and attraction organizers need to provide more services to accommodate diverse needs. Tourists from different nations conceivably have different language needs and expectations; tourists from different income levels are more likely to be interested in a broader range of activities. Third, tourists from different nations mean that more marketing investments have been made in foreign countries. This may also boost the CPI,

especially for the tourism-related industry. Fourth, tourists interacting with people from other nations and others with different income levels can learn from the activities others perform. This social learning process diversifies their travel experience and provides a new source of information. They may therefore make more purchases, leading to a higher CPI.

We expect future research to provide micro-foundations to explain the forces underlying these socio-demographic diversity measures on tourists and CPI measures in a tourism country. Second, the data used in this paper does not allow for the establishment of causality. Further research can explore the causal relationships between social structure and inflation (deflation) using observational causal analysis, which is useful in policy designs for economic development. Lastly, our results present interesting patterns in a tourism country. As more data become available, it would be interesting to see whether similar patterns persist comparing with other countries with a different economic structure.

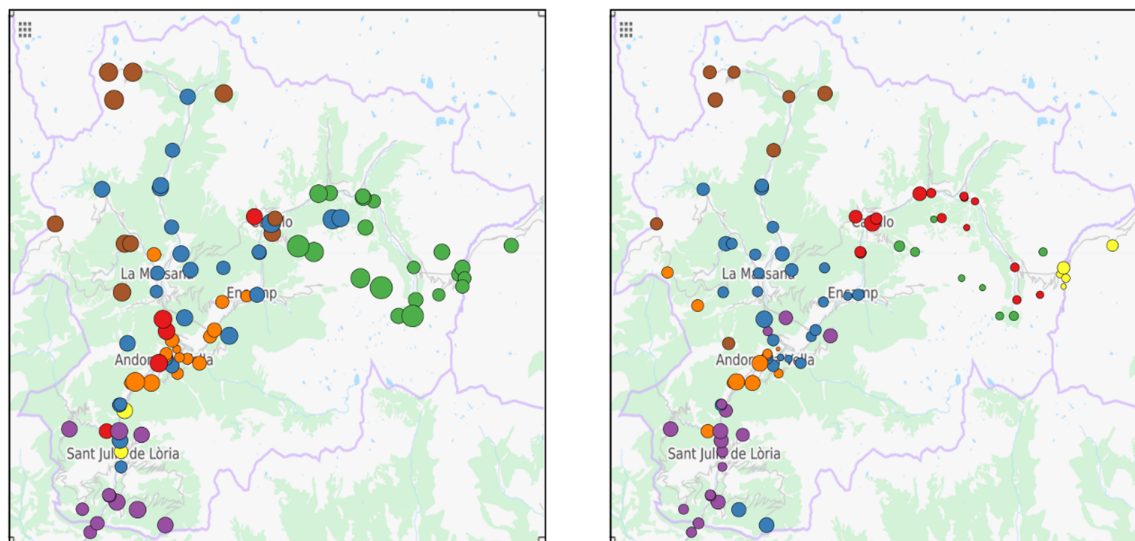


Fig. 8 Mapping general CPI in Andorra for general CPI in June, 2015 (left panel) and January, 2016 (right panel). The size of the nodes corresponds to the predicted CPI of the corresponding month. The color of the nodes corresponds to the segmentation based on the predicted CPIs across our observational period. We compute the pairwise correlations of cell towers and then perform spectral clustering on the correlation matrix to obtain the clusters.

Appendix: data availability

The data was obtained through a collaboration with the Andorra government and the Andorra mobile carriers. The brand of the phone and the country that the SIM card was registered was collected by the mobile carrier for billing purposes. The data is fully anonymized (no user identification is collected) and the data is stored on Andorran servers. The authors run algorithms on the Andorran server and obtain the aggregate statistics at cell tower levels (e.g., diversity measures used in this study).

Next, we will discuss the availability of such data to replicate the analysis of this research. We agree that the availability of mobile phone data and the practical applications of the method to assess the impacts of our paper. First, the application of mobile phone data into smart cities has been an active research field. In these applications, policy-making departments (e.g., transportation, tourism environment) can easily collaborate with the mobile carriers to utilize these data for social issues. The tourism and transportation department is a governmental organization and can typically collaborate with mobile carriers to obtain relevant data. These three organizations can collaborate collectively to build smart cities to improve and economy, similar to the type of collaboration built in this study. Second, in the case where the collaboration has not been established, mobile phone data has been made available by different stakeholders, specifically, open data by some mobile carriers and services (data and analytics) from some companies. Let us name a few of them. Owing to the wide-availability and the opportunistic nature (i.e., data initially collected for billing purposes) of mobile phone data, many mobile carriers share their data with the public. For example, in Europe, the CDR data has been made available in Milan and the Province of Trentino in Italy. In Africa, CDRs have been opened to the public in Ivory Coast and Senegal through two Orange D4D challenges. In Asian, China Unicom has shared the mobile phone data in 2018 CCF big data and computer intelligence competition through Data Foundation⁸. Except for these open data, some other organizations and companies provide services and analytics on mobile phone data. For example, AirSage provides the service in the US for collecting this type of CDR data that commercial companies can be used for identifying consumer patterns⁹.

Talkingdata is a China-based company that provides data and services on mobile phone data¹⁰. Predicio is based in Europe that provides mobile phone data and helps businesses with actionable consumer behavior insights¹¹. Flowminder is a Sweden-based company that provides anonymous phone records for policy-making and social good¹². Moreover, other firms provide a similar type of mobile phone records with higher spatial resolution through GPS, such as SafeGraph¹³ and Cubiq¹⁴. To allow the collective efforts to fight against COVID-19, SafeGraph has created a data consortium and shared mobility data in the US.

Data availability

Due to the nature of this research, data stakeholders did not agree for their data to be shared publicly, so supporting data is not available.

Received: 9 July 2020; Accepted: 1 June 2021;

Published online: 28 June 2021

Notes

- 1 A point of interest is a specific location that someone may find useful or attractive, such as a skiing resort or a museum. This term widely appears in geographic information systems. The category of POI potentially indicates the trip purposes and activities in transport studies (Leng et al., 2021).
- 2 Govern D'Andorra.
- 3 Note that CPI is a measure of the average *change* overtime in the prices paid by urban consumers for a market basket of consumer goods and services. Hence, it is usually reported relative to the past.
- 4 residence related services include (1) rental of housing, (2) services and products for the conservation of the home, (3) water distribution sewers and purification, (4) electric energy, (5) gas, flammable liquids.
- 5 This category includes: (1) furniture, furniture accessories, carpets, (2) textile articles for home and articles of furniture, (3) equipment and accessories for the home, (4) crystal, crockery and other home products, (5) small tools and disposable items for construction.
- 6 Nowcasting is the prediction of the present, the very near future and the very recent past in economics (Giannone et al., 2008).
- 7 When tuning the parameters for elastic net regression, the search space used for λ , the penalty term, was 10^{-3} to 10^3 . The possible range for values of α , the mixing

parameter between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$) regression, is 0 to 1. The optimal parameters (λ , α) found for each CPI category were: (0.419, 0.105) for general CPI, (0.296, 0.421) for hotels and restaurants, (0.470, 0) for food, drinks and tobacco, (0.944, 0.0) for clothes and shoes, (0.117, 0.895) for transport, (0.117, 0.053) for residence-related services, (1.501, 0.105) for furniture, home products and services, and (0.052, 0.053) for health.

8 <https://www.datafountain.cn/>.

9 <https://www.airsage.com>.

10 <http://mi.talkingdata.com>.

11 <http://www.predic.io>.

12 <https://web.flowminder.org>.

13 <https://www.safegraph.com>.

14 <https://www.cuebiq.com>.

References

- Alesina A, La Ferrara E (2000) Participation in heterogeneous communities. *Quar J Econ* 115:847–904
- Alesina A, Ferrara EL (2005) Ethnic diversity and economic performance. *J Econ Lit* 43:762–800
- AlShebli BK, Rahwan T, Woon WL (2018) The preeminence of ethnic diversity in scientific collaboration. *Nat Commun* 9:1–10
- Bakker MA et al. (2019) Measuring fine-grained multidimensional integration using mobile phone metadata: the case of syrian refugees in turkey. In: *Guide to mobile data analytics in refugee scenarios*. Springer, pp. 123–140.
- Bañbura M, Modugno M (2014) Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *J Appl Economet* 29:133–160
- Blumenstock JE (2016) Fighting poverty with data. *Science* 353:753–754
- Bok B, Caratelli D, Giannone D, Sbordon AM, Tambalotti A (2018) Macroeconomic nowcasting and forecasting with big data. *Ann Rev Econ* 10:615–643
- Cavallo A, Rigobon R (2016) The billion prices project: using online prices for measurement and research. *J Econ Perspect* 30:151–178
- Cia.gov. (2012) Cia world factbook entry: Andorra. Cia.gov
- Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328:1029–1031
- Fortune S (1987) A sweepline algorithm for voronoi diagrams. *Algorithmica* 2:153
- Galinsky AD et al. (2015) Maximizing the gains and minimizing the pains of diversity: a policy perspective. *Perspect Psychol Sci* 10:742–748
- Giannone D, Reichlin L, Small D (2008) Nowcasting: the real-time informational content of macroeconomic data. *J Monet Econ* 55:665–676
- Herring C (2009) Does diversity pay?: race, gender, and the business case for diversity. *Am Sociol Rev* 74:208–224
- Jean N et al. (2016) Combining satellite imagery and machine learning to predict poverty. *Science* 353:790–794
- Kemp S (2019) Digital 2019: global digital overview. <https://datareportal.com/reports/digital-2019-global-digital-overview>
- Lazer D, Radford J (2017) Data ex machina: introduction to big data. *Ann Rev Sociol* 43:19–39
- Leng Y, Rudolph L, Zhao J, Koutsopolous HN (2017) Synergistic data-driven travel demand management based on phone records. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*
- Leng, Y. et al. (2016) Urban computing using call detail records: mobility pattern mining, next-location prediction and location recommendation. Ph.D. thesis, Massachusetts Institute of Technology
- Leng Y, Narrowing A, Pentland A (2021) Tourism Event Analytics with Mobile Phone Data. *Forthcoming in ACM/IMS Journal of Data Science*
- Levine SS et al. (2014) Ethnic diversity deflates price bubbles. *Proc Natl Acad Sci USA* 111:18524–18529
- Macpherson DA, Sirmans GS (2001) Neighborhood diversity and house-price appreciation. *J Real Estate Finance Econ* 22:81–97
- Maxmen A (2019) Can tracking people through phone-call data improve lives?, url = <https://www.nature.com/articles/d41586-019-01679-5>, urldate = 29 May, 2019
- Modugno M (2013) Now-casting inflation using high frequency data. *Int J Forecast* 29:664–675
- Montalvo JG, Reynal-Querol M (2005) Ethnic diversity and economic development. *J Dev Econ* 76:293–323
- Monteforte L, Moretti G (2013) Real-time forecasts of inflation: the role of financial variables. *J Forecast* 32:51–61

Oliver N et al. (2020) Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances*. 6:eabc0764 (2020)

Ottaviano GI, Peri G (2006) The economic value of cultural diversity: evidence from us cities. *J Econ Geogr* 6:9–44

Pokhriyal N, Jacques DC (2017) Combining disparate data sources for improved poverty prediction and mapping. *Proc Natl Acad Sci USA* 114: E9783–E9792

Puritty C et al. (2017) Without inclusion, diversity initiatives may not be enough. *Science* 357:1101–1102

Rinaldo A et al. (2012) Reassessment of the 2010–2011 haiti cholera outbreak and rainfall-driven multiseason projections. *Proc Natl Acad Sci USA* 109:6602–6607

Stigler GJ (1961) The economics of information. *J Polit Econ* 69:213–225

Stirling A (2007) A general framework for analysing diversity in science, technology and society. *J R Soc Interf* 4:707–719

Wesolowski A et al. (2012) Quantifying the impact of human mobility on malaria. *Science* 338:267–270

Wesolowski A et al. (2015) Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proc Natl Acad Sci USA* 112:11887–11892

Workforce diversity: a key to improve productivity. *Proc Econ Finance* 11, 76–85 (2014)

Wright P, Ferris SP, Hiller JS, Kroll M (1995) Competitiveness through management of diversity: effects on stock price valuation. *Acad Manag J* 38:272–287

Acknowledgements

The authors would like to thank Andorra Telecom for providing the data used in this study. Meanwhile, the authors would like to thank Kent Larson and Luis Alonso from MIT Media Lab for the helpful discussions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-021-00822-w>.

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021