EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Identifying latent activity behaviors and lifestyles using mobility data to describe urban dynamics

Yanni Yang[1,2], Alex Pentland[2] and Esteban Moro[2,3*]

*Correspondence:
esteban.moroegido@gmail.com
[2]Connection Science, Institute for Data Science and Society, Massachusetts Institute of Technology, Cambridge, MA, United States
[3]Grupo Interdisciplinar de Sistemas Complejos (GISC), Department of Mathematics, Universidad Carlos III de Madrid, Leganés, Madrid, Spain
Full list of author information is available at the end of the article

Abstract

Urbanization and its problems require an in-depth and comprehensive understanding of urban dynamics, especially the complex and diversified lifestyles in modern cities. Digitally acquired data can accurately capture complex human activity, but it lacks the interpretability of demographic data. In this paper, we study a privacy-enhanced dataset of the mobility visitation patterns of 1.2 million people to 1.1 million places in 11 metro areas in the U.S. to detect the latent mobility behaviors and lifestyles in the largest American cities. Despite the considerable complexity of mobility visitations, we found that lifestyles can be automatically decomposed into only 12 latent interpretable activity behaviors on how people combine shopping, eating, working, or using their free time. Rather than describing individuals with a single lifestyle, we find that city dwellers' behavior is a mixture of those behaviors. Those detected latent activity behaviors are equally present across cities and cannot be fully explained by main demographic features. Finally, we find those latent behaviors are associated with dynamics like experienced income segregation, transportation, or healthy behaviors in cities, even after controlling for demographic features. Our results signal the importance of complementing traditional census data with activity behaviors to understand urban dynamics.

Keywords: Mobility data; Lifestyles; Topic analysis; Non-negative matrix factorization; Segregation; Health Risk; Transportation; Census

1 Introduction

Cities are the main ground on which our society and culture develop today. Most of our current understanding of problems like transportation, mobility, inequality, gentrification, or even social participation is based on census or survey information, which is updated infrequently, contains only coarse-grain information, and is scattered across different agencies or institutions [1]. On the other hand, we now have the potential to complement official data with high-resolution updates on how people purchase, move, get a job, or interact by leveraging new sources of information from mobile data [2, 3], social media [4, 5], wifi networks [6, 7], phone apps [8, 9], and credit cards [10, 11]. Companies have been using this wealth of data in the past. They are currently able to micro-segment clients based on their demographics and their behavioral traits [12–14]. However, most cities are still using primary segments of census groups (residential areas, housing prices, gender, age, unemployment) or small behavioral surveys to map problems like inequality, gentrification, or transportation. This approach falls short of anticipating, monitoring, or forecasting the rapid and complex evolution of those problems in our cities. For example, the recent pandemic has highlighted the shortcomings of using outdated, non-integrated, and slow-processed data to manage and anticipate the spreading of COVID-19 and the special relevance of real-time, more granular, and high-frequency mobility data [15, 16].

In particular, people's mobility data has become more available thanks to the prevalence of location acquisition techniques and mobile phones, and it enables a new way to study and understand human behavior in cities. People's mobility behavior, e.g., the places they visit and their visiting frequency, can reflect people's lifestyle, understood as "the way in which a person or group lives" [13, 17, 18]. Given the importance of lifestyles to predict individual and a group of individual's behavior, they have been thoroughly explored mainly in marketing [13] but also in many fields from transportation, [19], health [20–22] to psychology and sociology [14, 23].

The study of activity patterns and detection of lifestyles of urban residents based on survey data has a long tradition, [23] but recent developments in data collection and analysis have allowed the unveiling of the high-dimensional, rapid-changing, and complex lifestyles in our cities [10, 17–19, 24–28]. Studies that try to detect those lifestyles from activity data are generally limited by the completeness of the activity/mobility space (only expenditure patterns [10], mobility patterns only when mobile phone calls and messages appear [29, 30], only transportation transit patterns [25], or a very small number of demographic variables [19]), the limited geography (only one or two cities [10, 26]), or the number of people used to detect lifestyles [17, 27]. As a result, a small number of meaningful lifestyles were detected, insufficient to accommodate the highly heterogeneous and complex variability of our cities' behaviors.

On the other hand, in those studies, city residents' behaviors are typically classified into a single lifestyle group [10, 19]. This forces us to divide very similar individual behaviors into different groups just based on slight differences. Consequently, a significant fraction of individuals end up with unclassified lifestyles groups [10], or across groups with minimal different characteristics [19]. These problems severely limit those lifestyle groups' potential applicability to understanding problems like social-economic integration, mobility, or health, since slight individual behavioral differences or even different or incomplete datasets can yield a different grouping of users or lifestyles [18, 31–33].

In this work, we uncover people's lifestyles using a dataset of mobility traces of more than 1.2 million anonymous, opted-in users in 11 cities in the United States. By formulating people's behavior using venue and temporal activity vectors, we can extract a set of interpretable latent activity behavioral patterns [34]. Those latent behaviors are groups of visitation patterns that frequently co-occur in our sample of users. People's lifestyle is not a label for each individual but rather a linear combination of those latent behaviors with different weights. We investigate whether those behaviors can be predicted by simple demographic traits, e.g., race, income, or transportation. Although we find a small correlation, latent behaviors seem to be primarily independent of those demographic traits. Finally, we find that each component of those latent patterns has a different relationship with social, mobility, and health problems. Our results indicate that it is possible to construct a *be-*

havioral rich census of lifestyles in the U.S. cities that can complement traditional census to understand the main processes and problems in our cities.

2 Methods

2.1 Mobility data

Our primary data source is from Cuebiq, a location intelligence, and measurement company that, in 2017 supplied six-month-long records of anonymized, privacy-enhanced, and high-resolution mobile location pings across 11 U.S. census core-based statistical areas (CBSAs), see Additional file 1, Supplementary Note 1. Cities are defined as the Census Core Based Statistical Areas [35] that are socioeconomically metropolitan areas related to an urban center. It consists of approximately 67 billion records from $N = 1.2$ million anonymous opted-in devices, each of which has reported a total of at least 2000 locations over the six-month observation period. Our second data source is a collection of approximately 1.1 million verified venues across all CBSAs, obtained via the Foursquare API in 2017. Those venues are classified into different categories according to the Foursquare Category Hierarchy [36]. Only users with more than 50 visits during the period and with at least 5 categories visited were considered. We only considered the top most visited 248 categories to prevent over-fitting to small, infrequent categories. We infer the *home area* of each individual at the Census Block Group level using their most common location between 10 pm and 6 am. We further extract any individual visits to a given place that lasts for more than 5 minutes (see Additional file 1, Supplementary Note 1) and are less than 4 hours long. We have also tested that our results do not depend on this definition of visits to POI, see Additional file 1, Supplementary Note 1. It is important to note that our visitation patterns include not only consumption patterns (restaurants, shops, sports events, etc.) but also other non-commercial activities (transportation, education, health, outdoor activities, etc.), which are important to explain urban lifestyles.

2.2 Demographic data

Due to the anonymous nature of our location data, we obtained the demographic characteristics of each user at the area level. Demographic data like median household income, the fraction of Black population, the fraction of people that use public transportation, or urban characteristics like population density were obtained from the Census 2013-2017 ACS 5-year Estimates [37]. The fraction of people with more than 45 minutes of commuting was also obtained from the ACS Data.

2.3 Social, transportation and health data

To assess the importance of the latent behaviors, we validate them using different social, transportation and health data. The source of transportation data is the 2017 Local Area Transportation Characteristics for Households Data done by the Bureau of Transportation Statistics (BTS) [38], while obesity prevalence and physical activity are given by the 500 Cities Project Data from the Center for Disease Control (CDC) [39]. Obesity prevalence is measured as the percentage of adults, aged 18 or older, who report a body mass index (BMI) of 30 or higher. Physical activity is measured as the fraction of adults who report getting leisure-time physical activity in the past month. Cities are defined as the Census Core Based Statistical Areas [35] that are socioeconomically metropolitan areas related to an urban center. Note that although we could have constructed the mobility

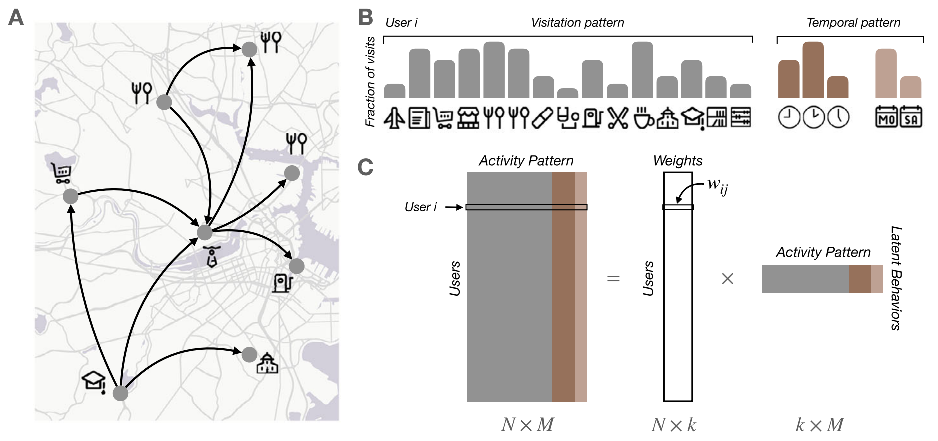

Figure 1 *Detecting latent behaviors* (**A**) Using individuals' trajectories, we identify the visits to the different places and the categories of those places. (**B**) Each individual is described by a *M*-dimensional (normalized) vector that contains the fraction of visits to each of the 286 categories (visitation pattern) plus the fraction of visits at different times during the day and the week (temporal pattern). (**C**) Non-negative matrix factorization is used to decompose the matrix of the *M*-dimensional vectors for each of our *N* users into a matrix of *k* different latent behaviors and the corresponding behavior's weights for each user. Icons designed by bqlqn/flaticon.com and Boston maps produced using Open Street Map data

variables using our data, we used the BTS and Census data because the latter is based on more reliable estimation statistics. But also because our data do not have complete daily individual trajectories of people, preventing us from having a precise estimation of the distance traveled and the commuting time.

To measure experienced social-economic integration, we use the inequality metric introduced in [8] to estimate how unequal is the exposure of an individual to the different income groups in the city. To this end, we divide the sample of users in each city into four quartiles according to the median household income of their home Census Block group [37]. Social-economic integration was measured as $I_i = 1 - \frac{2}{3}\sum_q |\tau_{iq} - 1/4|$, where τ_{iq} is the proportion of time user i is exposed to group q of income. That proportion is calculated by looking at the weighted distribution of income of the people that i encounters in the places she visits. Specifically $\tau_{iq} = \sum_\alpha \tau_{i\alpha}\tau_{q\alpha}$, where $\tau_{i\alpha}$ is the fraction of time that i spends at place α and $\tau_{q\alpha}$ represents the proportion of time at place α spent by income group q. Our metric for individual economic integration can be thought of as an extension of the traditional metric of isolation or interaction for groups to the level of individuals based on daily encounters among them. Finally, social exploration is measured as $E_i = S_i/N_i$, where S_i is the total number of different places visited by i and N_i is the total number of visits to places by i. See [8] for more details on these metrics and their distribution.

2.4 Non-negative matrix factorization

Rather than describing a person by a unique pattern, we will assume that there are some latent behavioral patterns that, when combined, define a person's lifestyle. The weight of the different latent behavior patterns could reflect their dominance over the person's lifestyle. To detect those latent behavioral patterns, we describe the activity of each user i by a M dimensional vector, which includes the (normalized) number of visits to the different types of places (248 venue categories, see Fig. 1 and Additional file 1, Supplementary Figure S2). We also include five temporal features about the fraction of those visits that happen during the morning (5 am to noon), afternoon (noon to 6 pm), evening/nighttime (6 pm to midnight), and also during the weekend and weekdays. Thus user i activity is described by a vector $\mathbf{x}_i$ of $M = 248 + 5$ components. As we will see in the results, adding the temporal pattern allows to recover different latent behaviors by day of the week.

When put together for all N users, they form a matrix X of $N \times M$ dimensions. Different methods exist to learn the latent patterns for the vectors of those N users, like spectral methods [3], Latent Dirichlet Allocation [25, 27, 29], neural networks, [17] or complex networks [10]. In general, these methods detect latent patterns, i.e. co-occurrence of variables (visitations in our case) that frequently appear in the dataset. Given that our vectors are non-negative, we apply non-negative matrix factorization (NMF) to X. NMF is a powerful technique for finding parts-based, linear representations of non-negative data and has been applied successfully in several applications like genomics, image recognition, or text mining [40, 41]. In the context of human behavior, it has also been used to identify the activity patterns of users or urban areas [26, 30, 42–44].

The key idea is that the activity matrix can be decomposed into two matrices $X = W \cdot B$, where B is a matrix of dimension $k \times M$ that contains each of the k latent behaviors and W is a $N \times k$ matrix that contains the weights of those latent behaviors for each user (see Fig. 1(C)). Thus each $\mathbf{x}_i$ can be decomposed as linear combination of k latent behaviors $\mathbf{x}_i = \sum_{j=1}^{k} w_{ij}\mathbf{b}_j$. In each latent behavior pattern $\mathbf{b}_j$, the higher value the type of venue or temporal feature holds, the more dominant that category of place or time of the week acts in the latent behavior pattern. For each user i, the higher the w_{ij}, the more important is latent behavior $\mathbf{b}_j$ to explain her activity.

The activity matrix X is factorized using different ranks k. To do that, we used non-negative matrix factorization (NMF) using fast sequential coordinate-wise descent and Kullback–Leibler (KL) divergence for the loss function. We run the NMF two hundred times for each value of k. Different methods [30, 44] were used to assess the value of k (see Additional file 1, Supplementary Note 4), including bi-cross validation [45]. We found that $k = 12$ was the one that optimizes the error, and the stability of the weights W and the bi-cross validation error, while making the latent behaviors B more interpretable across realizations. For that $k = 12$, we chose B and W from the realization with smaller KL loss (see Additional file 1, Supplementary Note 4). Our NMF factorization also produces very close representations of the whole mobility data. As we can see in Additional file 1, Supplementary Note 4, the KL average distance $d_{KL}(X, W \cdot B) = 0.0195 \pm 0.0001$ or the Frobenius distance $d_F(X, W \cdot B) = 0.0033 \pm 0.00004$ between the original data and the reconstructed one are very small. This means that on average, the error of our approximation is below 15% relative error.

To prevent an over-representation of the larger cities in the factorization, we did not use our 1.2 million users in the NMF. Instead, we randomly selected 10k users in each city and constructed the matrix X_0. We factorize it into $W_0 \cdot B_0$ and use B_0 to solve the non-negative linear regression problem $X \sim W \cdot B_0$ to get W for the rest of the users. This way, we get a fair representation of all the latent behaviors commonly present across cities. In any case, we have also checked that our results are robust against different definitions of the sample of users, see Additional file 1, Supplementary Note 7. To compare with other methods to detect latent patterns, we have also used LDA to detect latent behavior patterns (see Additional file 1, Supplementary Note 5). Although the results are somehow similar, we

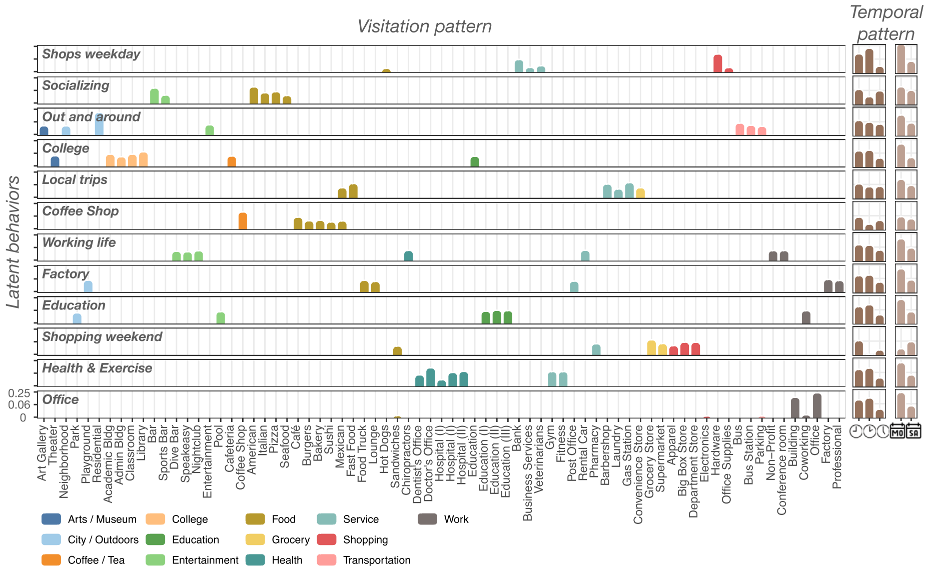

Figure 2 *Latent behaviors:* Visitation and temporal components for each of the $k = 12$ latent behaviors detected. For simplicity, only the top 7 venue category components by behavior are shown. Colors correspond to the different classifications of the venues. Temporal patterns correspond to the fraction of morning, afternoon, and night visits together with the fraction of weekday and weekend visits. The "Out and around" main components are residential and bus transportation during weekdays, while "Socializing" is composed of visitations to bars, sports bars, and food places like Italian, American, pizza, or seafood, especially during the evenings/nights. Finally "Local Trips" refers to a latent behavior mainly composed of running errands (laundry, gas station, convenience store) and visitation to fast food or food truck venues

find that the latent behavior patterns detected by NMF have better interpretations than those from LDA.

3 Results

3.1 Latent behaviors

The moderately large number of latent behaviors shows the richness and heterogeneity of our dataset. To interpret the latent behavior patterns, we first look into the different categories' dominance values and time slots (see Fig. 2). As we can see, most of the latent behaviors are easily recognizable and, as expected [26], their most relevant components belong, generally speaking, to combinations of working, food, entertainment, or shopping activities. Note that they are not strict projections only on one of those dimensions. For example, we find a latent behavior ("Working life") of working-related activities (Conference Room, Non-Profit) and nightlife venues, or a latent behavior "Out and around" that combines public transportation (bus) with neighborhood visits. Our choice of including the daily and weekly temporal pattern allows us to detect even different shopping behaviors between weekends ("Shopping weekends" that also includes grocery shopping) and weekdays ("Shops weekdays"). Other distinct latent behaviors correspond to "College" students, "Coffee shop" frequenters, or "Health & Exercise" visitors. Note that our denomination of the latent behaviors is based on the most dominant categories, which are also the most visited categories in cities. This does not mean that other less-visited categories are not part of those behaviors. For example, most latent behaviors have some components in the Food category (see Additional file 1, Supplementary Figure S2). However, their relative importance is smaller than in the “Local trips”, “Coffee Shop” or “Bar + Food” latent behaviors. Nevertheless, even for the places that are visited less frequently in our cities, the NMF can detect distinct patterns there. For example, the “Coffee Shop” latent behavior has large components in airport transportation venues than the rest of the behaviors (see Additional file 1, Supplementary Figure S2). Finally, it is worth noticing that our detected latent behaviors are not only related to expenditure: an analysis based only on expenditure patterns would have probably missed important latent behaviors like “Out and around”, “Office”, “College” or “Education”.

By definition, our latent behaviors encode those visitation patterns which are more frequently happening together. For example, we find in the “Local trips” latent behavior that fast food consumption is related to local errands, while nightlife is associated with work-related places like Conference Rooms or Rental Cars. “Shopping weekend” tells us that people tend to bundle visits to pharmacies, retail, groceries or departmental stores together. At the same time, in “Coffee Shop,” we see that visits to coffee shops are associated with the consumption of some types of food like Bakery, Sushi, and Burgers. Thus, our latent behaviors also tell us how people organize their mobility visitation patterns in their daily lives.

As we said, each person’s activity lifestyle can be described as a linear combination of the latent behavior patterns according to their weights obtained by NMF (see Fig. 3). Instead of being described as a simple latent behavior, we find that a user’s lifestyle generally depends on many behavior patterns. This is shown by the large entropy of weights by user $S_i = -\sum_{j=1}^{k} \hat{w}_{ij} \log \hat{w}_{ij} / \log k$ (where $\hat{w}_{ij}$ is the normalized weight by user), see Fig. 3(c). Strict dominance of single latent behavior would make $S_i = 0$, while we get $S_i = 1$ if all latent behaviors are present and equally important. In our data, we obtained that the average entropy is $\overline{S}_i = 0.67 \pm 0.15$, and thus many latent behaviors configure each user’s activity pattern.

However, some latent behaviors are, in general, more frequent than others. For example we find that the average weight of “Shopping Weekend” ($\overline{w}_{ij} = 0.774 \pm 0.001$) or “Office” ($\overline{w}_{ij} = 0.657 \pm 0.001$) is larger than others like “Working life” ($\overline{w}_{ij} = 0.197 \pm 0.001$) or “College” ($\overline{w}_{ij} = 0.195 \pm 0.001$), which signals that, as expected, the former latent behaviors are more common in urban areas than the latter ones. That is, most individuals have a very small (or even zero) “College” latent behavior, while most people have some weight on “Shopping Weekend” behavior.

Interestingly, we find that these results are pretty robust across all the cities studied (see Fig. 3(B)), which means that each latent behavior’s relative weights are very similar despite the different geography, density, or even cultural nature of the cities. This is an important result that demonstrates our method’s robustness across different cities and the homogeneity of activity patterns across the U.S. However, there are small but important variations. For example, the relative weight of the “College” behavior is larger in Boston $\overline{w}_{ij}^{\text{Boston}} = 0.270 \pm 0.001$ because of the large population of university students in the area. Also, cities with better public transportation systems like Boston, Washington DC, or N.Y., have larger weights in the “Out and around” latent behavior than cities like Dallas or Detroit, where public transportation is scarce. Other significant variations happen in the “Coffee Shop” latent behavior, more present in cities like San Francisco or Seattle than in the rest. This is expected given that those cities are the ones with the most coffee shops per capita [46]. Taking together these results shows the adaptability and robustness of our

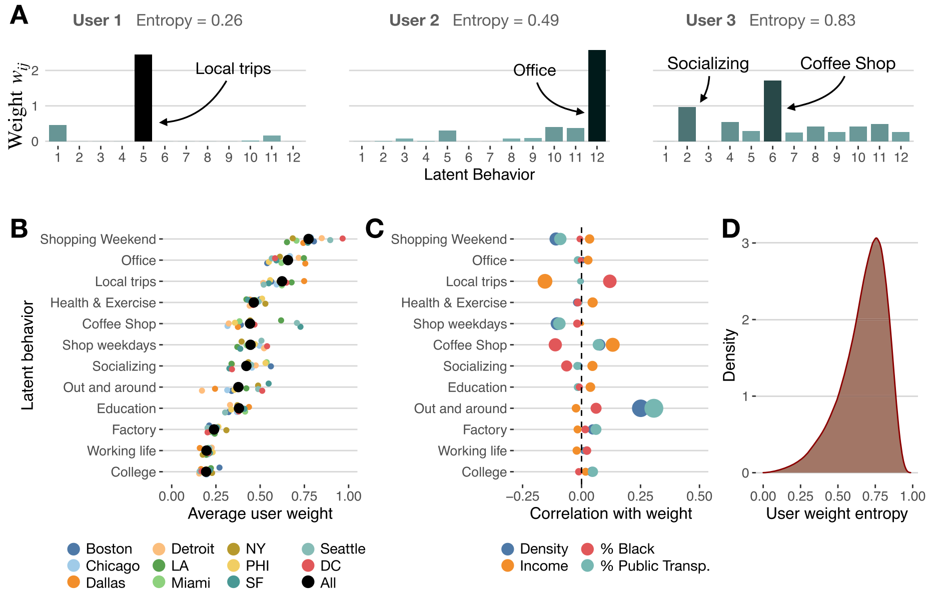

Figure 3 *Users' latent behaviors and lifestyles* (**A**) Distribution of weights by latent behavior for three different users in our database. User 1 behavior is dominated by the "Local trips" behavior and has low entropy. However, user 3 has a larger entropy, and her behavior is a very diverse mixture of latent behaviors, including "Bar + Food" and "Coffee Shop". (**B**) Average weight for the different latent behaviors in all areas (black) and different cities. (**C**) Correlation between the weight of latent behaviors and different demographic and urban characteristic features like population density, median income, percentage of Black population, and percentage of people using public transportation. (**D**) Distribution of user's weight entropy

latent behaviors to describe the similarities and peculiarities of activity patterns in the different cities in the U.S.

3.2 Latent behaviors are not fully described by demographics or urban characteristics

One important question is whether the detected latent behaviors can be explained by individuals' simple demographic or urban characteristics. Group segmentation using census-type only data is traditional in marketing, [12, 47] and even the U.S. Census Bureau used it to design its 2020 campaign [48]. Since mobility is highly influenced by socioeconomic status, access to public transportation, or the density of the urban area, we could expect that the detected latent behaviors might depend strongly on those features. However, we find that users' weights w_{ij} are largely independent of the median income, population density, the fraction of Black population, or even the fraction of people using public transportation (see Fig. 3(C) and Additional file 1, Supplementary Table S1). We only find moderate correlations with some latent behaviors. For example, low-income people have more "Local trips" latent behavior, while "Coffee Shop" is a behavior more likely to be found in high-income areas. Of course, "Out and around" latent behavior is more prevalent in areas with higher public transportation use. Apart from those cases, the correlation between our latent behaviors' weights and demographic and urban features is very small $R^2 \leq 0.1$ (see Additional file 1, Supplementary Note 6 and Additional file 1, Supplementary Table S1). Thus, latent behaviors detected using mobility data are different from the traditional census demographic traits. The detected activity patterns give a different and complementary perspective of our cities than traditional census analysis, allowing us to construct a richer *behavioral census* that includes those behaviors.

3.3 Association of latent behaviors with social, mobility, and health problems

To demonstrate the complementary power of the latent behaviors to traditional census approaches, we have analyzed their association with different social, mobility, and health outcomes in the 11 cities. In the social dimension, we have considered I_i, the income integration (or diversity) experienced by each individual, introduced in [8]. This quantity reflects how homogeneous is the exposure of each individual to the different income groups in the city: by using the household median income of the Census Block group where user i lives, we can quantify the income group (income quartile within each city) she belongs to. Using that information for each user, we can estimate the amount of time a user i is exposed to the different income groups in the city while visiting different venues: if $I_i = 0$, individual i only goes to places where her particular income group is the majority. If $I_i = 1$, the individual is exposed equally to people from all the city's income groups (Methods Sect. 2.3). Other versions of diversity exposure have been analyzed recently [9] and, in particular, income exposure diversity is related to social capital and impacts economic opportunities, and social income mobility of individuals [49]. Also, along the social dimension, we have studied the individual's place exploration E_i, which measures the rate of visitation to different places by i in our time period [50]. That is, if N_i is the number of visits made by user i and S_i is the number of unique places visited, then $E_i = S_i/N_i$. Although people spend most of their time in a very small number of places [2, 24], it is well known that some tend to visit more places (*explorers*, $E_i \simeq 1$), while some others spend most of their time in a small set of places (*returners*, $E_i \simeq 0$). Explorers are people that go very often to never visited before places, while returners are constantly coming back to places that they already visited. Those two distinct classes of individuals have been found in many different mobility studies [2, 8, 51], but also in other people's activities like social network connections [52], web browsing [53], or knowledge discovery [54].

Both income exposure diversity and place exploration are crucial to understanding the social component of mobility in our cities and, specifically, how segregated (not integrated) people are. As was found in [8] experienced income integration is moderately and positively related to place exploration ($\rho = 0.456 \pm 0.001$). To test the association of the latent behaviors in these problems, we have used a regression model:

$$I_i, E_i \sim \sum_{j=1}^{k} \beta_j w_{ij} + \sum_{l=1}^{m} \gamma_l d_l + \mathrm{MSA}_i + \varepsilon_i, \tag{1}$$

where d_l refers to the four demographic and urban features mentioned before (median household income, the density of the area, the fraction of Black people, and the fraction of use of public transportation), and MSA_i is a fixed factor by city (Metropolitan Statistical Area). Including the census variables and city-fixed effects allows us to investigate the fundamental role of latent behaviors once we control for potential effects by demographic and urban characteristics and the city where users live. In our models, census features are always less important to explain that variability ($R^2 = 0.059$ for I_i and $R^2 = 0.025$ for E_i only using census variables) than our latent behaviors ($R^2 = 0.164$ and $R^2 = 0.26$ respectively using also latent behaviors, see Additional file 1, Supplementary Tables S2 and S3. This result shows that our latent behaviors encode most of the social-economic integration and exploration variability across users, which are largely independent of census variables.

Nevertheless, not all latent behaviors have the same effect on experienced income segregation and exploration. Figure 4 shows the relationship [measured as the standardized coefficient β_j in the model (1)] of the different latent behaviors with both social problems. We find that the latent behaviors that impact economic integration are related to shopping or food/coffee. In contrast, others like college, office, or health are not heavily associated with economic integration. Interestingly, behaviors like "Shopping Weekend" or "Coffee Shop" are positively associated with economic integration, while behaviors "Education", "Factory" or "Shopping weekdays" are more present in users with more considerable experienced income segregation. This might be explained by the fact that shops, coffee shops, or some restaurants are more economically diverse in the city than factories, education, or local shops [8]. In general social exploration follows the same pattern, although people with more "Local trips" behavior tend to be more explorers without being more integrated. These results show that the latent behaviors carry significant explanatory power of the income diversity and exploration experienced by people in the urban areas analyzed.

We have also studied other problems related to transportation and health. Lifestyles are crucial to understanding our mobility choices or opportunities and transportation routines [19], but also physical activity and the prevalence of some health conditions [21]. Since we do not have individual health conditions, we have used data from the Census, the Bureau of Transportation Statistics (BTS), and the Center for Disease Control to take variables by census tract α describing the average distance traveled by residents D_α, the fraction of people that have more than 45 minutes of commuting C_α, the fraction of people with leisure-time physical activity in the past month P_α and the fraction of people with no obesity O_α (see Methods Sect. 2.3). To study the relationship of our latent behaviors with those problems, we construct the average weight by census tract $\hat{w}_{\alpha,j}$ for all the users living in that tract α and fit them using similar models as Eq. (1), see Additional file 1, Supplementary Note 6.

As we can see in Fig. 4, local behaviors like "Out and around" have a strong negative association with the distance traveled, although the association with commuting duration is positive. In general, tracts with more shopping behavior travel more, while those with larger "Coffee Shop" or "Bar + Food" latent behavior have smaller commutes and distance traveled. Since our data is projected at the level of the census tract, individual variability is averaged out, and thus, demographic and urban variables are more important to explain the variability in transportation variables. However, our latent behaviors still explain part of how people commute or move around the city, even if we condition on income, race composition, density, or use of public transportation (see Additional file 1, Supplementary Note 6).

Finally, we see in Fig. 4 that, although not all of them are relevant, there are a fraction of latent behaviors that have a significant association with health outcomes. As expected, more presence of latent behaviors like "Local trips" (which include visits to Fast Food venues) is associated with less Physical Activity and more Obesity, while behaviors like "Out and around" or "Coffee Shop" are positively associated with the amount of Physical Activity and the absence of Obesity. However, social and mobility important behaviors like shopping are not significantly related to health outcomes. This is even after controlling

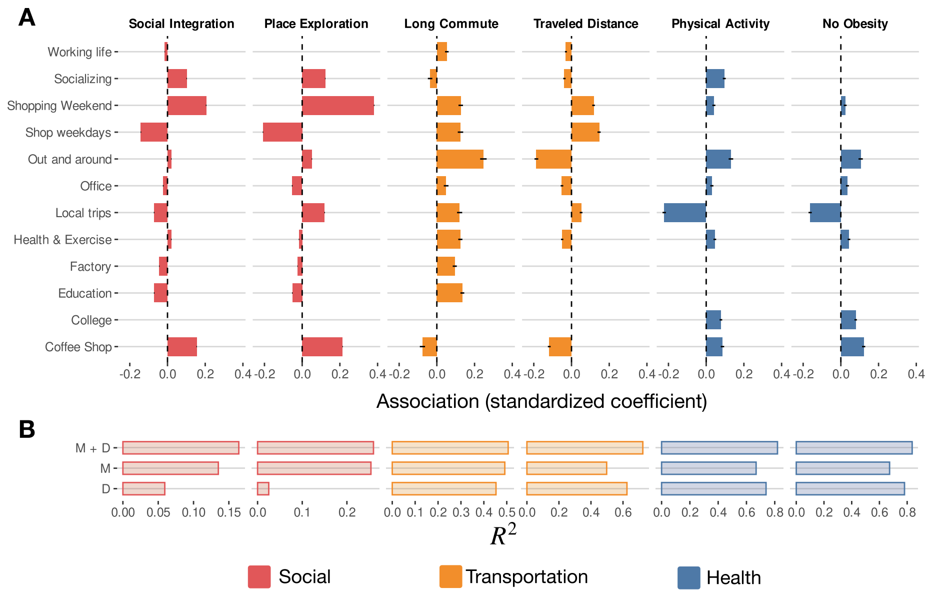

Figure 4 *Relationship of behavioral patterns with social, mobility, and health problems.* (**A**) Bars show the coefficient of each behavior weight (columns) on the different regression models for each social-economic integration, place exploration, the fraction of people with longer than 45 min commute, amount of distance traveled by day, the fraction of people doing daily physical activity, and the fraction of people not having obesity (rows). (**B**) Bar shows the R^2 for each model, including only the demographic variables D, only the mobility behavioral weights M, and both together $M + D$. Panel (**A**) correspond to the coefficients for the $M + D$ case. See Additional file 1, Supplementary Note 6 for the complete regression tables

for demographic variables like income or race, which are the most critical determinants of those health outcomes [55].

In summary, our results show that our latent behaviors that constitute individual lifestyles are significantly correlated with social, transportation, and health outcomes. Different behaviors are related to different dimensions, and in some cases, the importance of latent behaviors is similar to or even surpasses that of demographic variables. For example, knowing that a particular user has extensive shopping or food/coffee behaviors can better explain her economic integration and exploration than knowing her income (see Additional file 1, Supplementary Tables S2 and S3). Or conditioning on income or population density, we see that some behaviors like "Out and around" or "Coffee Shop" are associated with transportation and health outcomes.

4 Discussion

Understanding urban problems require a good description of human behavior in cities [1, 56]. Our research shows that the high-dimensional nature of mobility of millions of people visiting millions of places in the U.S. can be projected onto a small set of latent behaviors that capture their routines and habits that result from their choices or opportunities accessible to them. Demographics or urban characteristics cannot fully explain those latent behaviors, and people living in the same neighborhood with the same income, race, or educational levels might have different shopping, working, or leisure latent behaviors, resulting in entirely different lifestyles. Since the composition of the lifestyles is robust across different geographical areas and cities, our results could be used to build characterizations and compare individuals and groups at different geographical and demographic levels. This could enrich the current Census by including the composition of the different latent behaviors to study urban areas. It could also help in methods of exploiting mobility data by preserving the privacy of individuals. This can be done by computing projected aggregated variables along those latent behaviors rather than detailed and more invasive individual visitation patterns.

Our latent behaviors describe how people organize their visitation patterns and mobility around the city. For example, we find that people that make trips to errands also visit fast food outlets frequently ("Local trips" latent behavior), while heavy users of public transportation (bus) also spend much time in the neighborhood and entertainment ("Out and around"). Working life is also related to nightlife. These dependencies show that those latent behaviors represent combined aspects of our life that occur concurrently and which could be used to devise successful holistic interventions to change people's lifestyles. For example, people that run many errands might choose fast food because they are time-poor or because errands take place around specific food environments (food swamps). Our results can help design public health interventions that incorporate those distinct lifestyles to identify those routines and habits that are most risky for health [57].

We note that latent behaviors have a different relationship with social, transportation, and health outcomes. For example, while weekend shopping behaviors are associated with more exposure to economic diversity of urban dwellers, they carry more commuting time and travel distance and, thus, more pollution. Similarly, the "Out and around" latent behavior is associated with longer commutes and more physical activity or the absence of obesity. Since most urban interventions are likely to change the relative weight of those latent behaviors or ultimately change them completely, it is essential to balance the trade-off among social, transportation, and health outcomes encoded in those behaviors. Also, not all behaviors have the same weight in describing users' lifestyles. Shopping, food, or working latent behaviors are the most important, suggesting that they are the ones where interventions to change experienced income segregation, transportation, or health outcomes could be more considerable [25].

Our results show that activity lifestyles are not monolithic groups of homogeneous behavior among people. Our framework of describing lifestyles as a combination of latent behaviors reflects that lifestyles are instead a continuum spectrum of the relative balance between work, shopping, transportation, or leisure time. Given the ubiquitous nature of mobility and activity data from mobile phones, we hope this framework could be used in the future to understand better the rapid and extensive scale changes in other urban areas and cities worldwide.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00390-w.

Additional file 1. Additional file contains Supplementary Note 1—Data, Supplementary Note 2—Representativity, Supplementary Note 3—Non-negative matrix factorization, Supplementary Note 4—Rank Selection, Supplementary Note 5—Comparison with LDA, Supplementary Note 6—Models, and Supplementary Note 7—Robustness checks. It also contains Supplementary Figures S1 to S5 and Supplementary Tables S1 to S4. (PDF 838 kB)

Acknowledgements

We would like to thank Cuebiq for granting us access to mobility data through their Data for Good program. We would like to thank the MIT ILP program and Ferrovial for their financial support for this project. E.M. acknowledges support by

Ministerio de Ciencia e Innovación/Agencia Española de Investigación (MCIN/AEI/10.13039/501100011033) through grants FIS2016-78904-C3-3-P and PID2019-106811GB-C32, and the National Science Foundation under Grant No. 2218748.

Availability of data and materials

The data that support the findings of this study are available from Cuebiq through their Data for Good program, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Aggregated data used in the models are however available from the authors upon reasonable request and with permission of Cuebiq. Custom code that supports the findings of this study is available from the corresponding author upon request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contributions

All authors designed research. YY and EM performed all the calculations and analysis. All the authors contributed to the writing of the article. All authors read and approved the final manuscript.

Author details

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. [2]Connection Science, Institute for Data Science and Society, Massachusetts Institute of Technology, Cambridge, MA, United States. [3]Grupo Interdisciplinar de Sistemas Complejos (GISC), Department of Mathematics, Universidad Carlos III de Madrid, Leganés, Madrid, Spain.

Received: 9 August 2022 Accepted: 9 May 2023 Published online: 18 May 2023

References

1. Cagney KA, Cornwell EY, Goldman AW, Cai L (2020) Urban mobility and activity space. Annu Rev Sociol 46(1):1–26
2. Song C, Koren T, Wang P, Barabasi A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823
3. Eagle N, Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63(7):1057–1066
4. Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. PLoS ONE 10(5):0128692
5. Huang Q, Wong DW (2016) Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? Int J Geogr Inf Sci 30(9):1873–1898
6. Kontokosta CE, Johnson N (2017) Urban phenology: toward a real-time census of the city using wi-fi data. Comput Environ Urban Syst 64:144–153
7. Bellini P, Cenni D, Nesi P, Paoli I (2017) Wi-fi based city users' behaviour analysis for smart city. J Vis Lang Comput 42:31–45
8. Moro E, Calacci D, Dong X, Pentland A (2021) Mobility patterns are associated with experienced income segregation in large US cities. Nat Commun 12(1):4633
9. Athey S, Blei D, Donnelly R, Ruiz F, Schmidt T (2018) Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. AEA Pap Proc 108:64–67
10. Di Clemente R, Luengo-Oroz M, Travizano M, Xu S, Vaitla B, Gonzalez MC (2018) Sequences of purchases in credit card data reveal lifestyles in urban populations. Nat Commun 9(1):3330
11. Krumme C, Llorente A, Cebrian M, Pentland AS, Moro E (2013) The predictability of consumer visitation patterns. Sci Rep 3(1):1645
12. 2018 Esri Tapestry Segmentation Methodology. https://support.esri.com/en/white-paper/3575. Accessed: 14-12-2020 (2018)
13. Mitchell A (1983) The nine American lifestyles: who we are and where we're going. Scribner Book Company, New York
14. Kahle LR, Beatty SE, Homer P (1986) Alternative measurement approaches to consumer values: the list of values (lov) and values and life style (vals). J Consum Res 13(3):405
15. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, Nadai MD, Letouzé E, Salah AA, Benjamins R, Cattuto C, Colizza V, Cordes N, Fraiberger SP, Koebe T, Lehmann S, Murillo J, Pentland A, Pham PN, Pivetta F, Saramäki J, Scarpino SV, Tizzoni M, Verhulst S, Vinck P (2020) Mobile phone data for informing public health actions across the Covid-19 pandemic life cycle. Sci Adv 6(23):0764
16. Aleta A, Martín-Corral D, Piontti AP, Ajelli M, Litvinova M, Chinazzi M, Dean NE, Halloran ME, Longini IM Jr, Merler S, Pentland A, Vespignani A, Moro E, Moreno Y (2020) Modelling the impact of testing, contact tracing and household quarantine on second waves of Covid-19. Nat Hum Behav 4:964–971
17. Zion EB, Lerner B (2018) Identifying and predicting social lifestyles in people's trajectories by neural networks. EPJ Data Sci 7(1):45
18. Kitamura R (2009) Life-style and travel demand. Transportation 36(6):679–710
19. Salomon I, Ben-Akiva M (1983) The use of the life-style concept in travel demand models. Environ Plann A, Econ Space 15(5):623–638
20. Sadilek A, Kautz H (2013) Modeling the impact of lifestyle on health at scale. In: Proceedings of the sixth ACM international conference on web search and data mining, pp 637–646
21. Joumard I, Andre C, Nicq C, Chatal O (2010) Health Status Determinants: Lifestyle, Environment, Health Care Resources and Efficiency. SSRN Electron J
22. Matz CJ, Stieb DM, Brion O (2015) Urban-rural differences in daily time-activity patterns, occupational activity and housing characteristics. Environ Health 14(1):1–11

23. Hanson S, Hanson P (1981) The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. Econ Geogr 57(4):332–347
24. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782
25. Zhao Z, Koutsopoulos HN, Zhao J (2020) Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model. Transp Res, Part C, Emerg Technol 116:102627
26. Hu T, Bigelow E, Luo J, Kautz H (2016) Tales of two cities: using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. IEEE Trans Big Data 3(1):55–66
27. Farrahi K, Gatica-Perez D (2011) Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans Intell Syst Technol 2(1):1–27
28. Ma J, Li B, Mostafavi A (2022) Characterizing Urban Lifestyle Signatures Using Motif Properties in Network of Places. arXiv preprint. arXiv:2204.01103
29. Xu S, Di Clemente R, González MC (2019) Mining urban lifestyles: urban computing, human behavior and recommender systems. arXiv preprint. arXiv:1911.05464
30. Aledavood T, Kivimäki I, Lehmann S, Saramäki J (2022) Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data. Sci Rep 12(1):5544
31. Hill JO (2009) Can a small-changes approach help address the obesity epidemic? A report of the joint task force of the American society for nutrition, institute of food technologists, and international food information council. Am J Clin Nutr 89(2):477–484
32. Jiang S, Ferreira J, González MC (2012) Clustering daily patterns of human activities in the city. Data Min Knowl Discov 25(3):478–510
33. Xi W, Calder CA, Browning CR (2020) Beyond activity space: detecting communities in ecological networks. Ann Am Assoc Geogr 110(6):1787–1806
34. Toch E, Lerner B, Ben-Zion E, Ben-Gal I (2019) Analyzing large-scale human mobility data: a survey of machine learning methods and applications. Knowl Inf Syst 58(3):501–523
35. United States Census Bureau. Core-Based Statistical Areas. https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html. Accessed: 22-06-2019 (2000)
36. Foursquare Venue Category Hierarchy. https://developer.foursquare.com/docs/build-with-foursquare/categories/. Accessed: 09-12-2020 (2020)
37. United States Census Bureau: 2013-2017 American Community Survey 5-year Estimates. https://www.census.gov/programs-surveys/acs. Accessed: 2020-12-04 (2019)
38. Bureau of Transportation Statistics: Local Area Transportation Characteristics for Households Data. https://www.bts.dot.gov/latch/latch-data. Accessed: 08-01-2021 (2017)
39. Centers for Disease Control and Prevention: 500 Cities: local data for better health. https://www.cdc.gov/places/about/500-cities-2016-2019/index.html. Accessed: 08-01-2021 (2017)
40. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
41. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci 101(12):4164–4169
42. Gauvin L, Panisson A, Cattuto C (2014) Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. PLoS ONE 9(1):86028
43. Graells-Garrido E, Caro D, Parra D (2018) Toward Finding Latent Cities with Non-Negative Matrix Factorization. arXiv:1801.09093
44. Møllgaard PE, Lehmann S, Alessandretti L (2022) Understanding components of mobility during the Covid-19 pandemic. Philos Trans R Soc A 380(2214):20210118
45. Owen AB, Perry PO (2009) Bi-cross-validation of the SVD and the nonnegative matrix factorization. Ann Appl Stat 3(2):564–594
46. Apartment Guide: The Best Cities for Coffee Lovers in America. https://www.apartmentguide.com/blog/best-cities-for-coffee-lovers/. Accessed: 19-01-2021 (2020)
47. Spielman SE, Singleton A (2015) Studying neighborhoods using uncertain data from the American community survey: a contextual approach. Ann Assoc Am Geogr 105(5):1003–1025
48. Vines M, Bates N, Scheid S, Trejo YG (2019) 2020 Census Predictive Models and Audience Segmentation Report. https://www.census.gov/programs-surveys/decennial-census/2020-census/research-testing/communications-research/predictive-models-audience-segmentation-report.html
49. Chetty R, Jackson MO, Kuchler T, Stroebel J, Hendren N, Fluegge RB, Gong S, Gonzalez F, Grondin A, Jacob M, Johnston D, Koenen M, Laguna-Muggenburg E, Mudekereza F, Rutter T, Thor N, Townsend W, Zhang R, Bailey M, Barberá P, Bhole M, Wernerfelt N (2022) Social capital I: measurement and associations with economic mobility. Nature 608(7921):108–121
50. Alessandretti L, Sapieżyński P, Sekara V, Lehmann S, Baronchelli A (2018) Evidence for a conserved quantity in human mobility. Nat Hum Behav 2(7):485–491
51. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6(1):8166
52. Miritello G, Lara R, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. Sci Rep 3(1):1–7
53. Barbosa HS, de Lima Neto FB, Evsukoff A, Menezes R (2016) Returners and explorers dichotomy in web browsing behavior—a human mobility approach. In: Complex networks VII: proceedings of the 7th workshop on complex networks CompleNet 2016. Springer, Berlin, pp 173–184
54. Singh CK, Tupikina L, Lécuyer F, Starnini M, Santolini M (2023) Charting mobility patterns in the scientific knowledge landscape. arXiv preprint. arXiv:2302.13054
55. Ogden CL, Fakhouri TH, Carroll MD, Hales CM, Fryar CD, Li X, Freedman DS (2017) Prevalence of obesity among adults, by household income and education—United States, 2011–2014. Morb Mort Wkly Rep 66(50):1369–1373

56. Batty M (2013) Big data, smart cities and city planning. Dialogues Hum Geogr 3(3):274–279
57. Riet J, Sijtsema SJ, Dagevos H, Bruijn G-JD (2011) The importance of habits in eating behaviour. An overview and recommendations for future research. Appetite 57(3):585–596

Publisher's Note

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