



Prediction of employment and unemployment rates from Twitter daily rhythms in the US

Eszter Bokányi* , Zoltán Lábszki and Gábor Vattay

*Correspondence:
bokanyi@complex.elte.hu
Department of Physics of Complex
Systems, Eötvös Loránd University,
Pázmány Péter sétány 1/A,
Budapest, H-1117, Hungary

Abstract

By modeling macro-economical indicators using digital traces of human activities on mobile or social networks, we can provide important insights to processes previously assessed via paper-based surveys or polls only. We collected aggregated workday activity timelines of US counties from the normalized number of messages sent in each hour on the online social network Twitter. In this paper, we show how county employment and unemployment statistics are encoded in the daily rhythm of people by decomposing the activity timelines into a linear combination of two dominant patterns. The mixing ratio of these patterns defines a measure for each county, that correlates significantly with employment (0.46 ± 0.02) and unemployment rates (-0.34 ± 0.02). Thus, the two dominant activity patterns can be linked to rhythms signaling presence or lack of regular working hours of individuals. The analysis could provide policy makers a better insight into the processes governing employment, where problems could not only be identified based on the number of officially registered unemployed, but also on the basis of the digital footprints people leave on different platforms.

Keywords: unemployment prediction; Twitter; social media; activity patterns

1 Introduction

Until recently, it has been a time-consuming, costly and arduous work to collect and analyze data about individual humans at a large scale. With the advent of the digital era, there is a growing amount of data accessible online that enables the analysis and modeling of human behavior. However, our understanding of these digital data sources and the methods that connect the data to real-world outcomes is still limited.

Several aspects on the possible usage of mobile phone records and social media status updates in the estimation of official data, such as census, demographic or land use records have been discussed in recent papers. A promising approach is the analysis of the diurnal rhythm of humans. Due to the 24 hour periodicity of the Earth's rotation, we are biologically bound to show daily periodic behavior both at the individual and at the aggregate level. This periodic cycle is governed mainly by internal biochemical processes [1–4], but the impact of external factors and the environment also leaves its imprint on these daily patterns [5, 6].

As Särämäki and Moro point out in their paper [7], an interesting application is to consider the geospatial aspects of the aggregate level of daily rhythms, as it can provide insight into several different phenomena ranging from the actual land use patterns in a city [8–18] and on a campus [10], to the tracking of anomalous events [18, 19], or the estimation of population size [20], mobility patterns [21], poverty [22] or crime rates [23] in a certain area.

Because these aggregate patterns always consist of the superposition of the daily rhythms of individuals, it is worth investigating how the main features of the aggregate level form from superposition. If we can cluster individuals into more or less homogeneously behaving groups based on their daily patterns [24], then the aggregate pattern can be understood as the combination of the group patterns, and the group that has more individuals dominates the aggregate daily rhythm. The groups of individuals can form along many demographic and/or socioeconomic factors, of which being employed and going to and from work at regular hours is the most determining one with respect to the daily activity patterns. Thus, decomposing the groups from the aggregate patterns in different geographical regions may give insight into the estimation of employment statistics in that region.

Nowcasting or estimating unemployment rates using the digital traces of search engines has already been in the focus of several papers [25–27]. It has already been shown, that daily activity patterns of individuals can be linked to the regularity of their working hours [28]. Because the loss of a job has severe psychological consequences [29], the effects of a mass layoff can be detected in the unemployment rates and provide a possibility of forecasting macro-economical effects based on observation of several individuals [30]. In [31], there is a strong evidence that aggregated daily activities of certain time intervals of geographical regions can be indicative of unemployment rates.

In this paper we obtain 63 million geolocated messages from the publicly available stream of the social network Twitter from the area of the United States sent between January and October 2014. We aggregate Monday to Friday relative tweeting activity for each hour in each US county to form an average workday activity pattern. We then assume that these activity patterns form a roughly linear subspace of the 24-hour “timespace”. By finding this linear subspace, that is, by finding the line on which the county patterns lie, we are able to give a measure that is linked to the ratio of two groups of people tweeting in a county. We then show that this measure correlates significantly with county employment and unemployment rates, and that the average patterns corresponding to the two groups can be linked to lifestyles connected to regular working hours or the lack of them. We thus give a possible framework for decomposing the digital activity patterns of geographical regions and linking the decomposition to employment and unemployment rates.

2 Methods

2.1 Twitter dataset

We use the data stream freely provided by Twitter through their Application Program Interface, which amounts to approximately 1% of all sent messages. In this study, we focus on the part of the data stream with geolocation information. These geolocated tweets originate from users who chose to allow their mobile phones to post the GPS coordinates along with a Twitter message. The total geolocated content was found to only comprise of a small percentage of all tweets; therefore with data collection focusing only on these,

a large fraction of all geotagged tweets can be gained [32]. Our dataset includes a total of 63 million tweets from the contiguous United States collected between January 2014 and October 2014. These are all geotagged - that is, they have GPS coordinates associated with them. We construct a geographically indexed database of these tweets, permitting the efficient analysis of regional features [33]. Using the Hierarchical Triangular Mesh scheme for practical geographic indexing [34, 35], we assigned a US county to each tweet. County borders are obtained from the GAdm database [36].

2.2 Demographic datasets

For the population-weighted linear model of the next section, we obtain county-level population statistics from the US 2010 Census [37]. We download the unemployment and labor force data for the time window of the Twitter dataset from the Local Area Unemployment Statistics page of the Bureau of Labor Statistics [38]. We take an average of the months ranging from January 2014 to October 2014 for each county.

Though unemployment levels are defined as the number of unemployed per total labor force in a county, we define the share of employed as the number of employed divided by the whole population of a county. This measure fits the model for the daily rhythm better as discussed in Section 3.

2.3 Daily activity patterns

We define a daily activity pattern with hourly resolution for each county, which are enumerated by $k = 1, \dots, M$. We take all tweets originating from a given county from the period between January 2014 and October 2014. Then we aggregate the number of tweets (n_i) in each hour (the hour range goes from $i = 0, 1, \dots, 23$) on workdays, that is from Monday to Friday, after correcting for timezone and daylight saving time in each county. Because of the differing population and Twitter penetration rates (share of people using Twitter) in each county, we normalize the number of tweets by the total number of tweets counted. Thus, each county (k) is represented by a 24-dimensional vector ($\mathbf{y}^{(k)}$), where the elements of $\mathbf{y}^{(k)}$ are:

$$y_i^{(k)} = \frac{n_i}{\sum_{i=0}^{23} n_i},$$

and obviously,

$$\sum_{i=0}^{23} y_i^{(k)} = 1 \quad \forall k = 1, \dots, M.$$

To improve the quality of our dataset, we consider only those counties in which the overall tweet count during the ten month exceeded the threshold of 1,800. Thus, we are left with 1,884 counties for our analysis.

2.4 Linear model

We assume that the tweeting pattern of a county can be represented by the linear combination of only two universal patterns (**A** and **B**) that are mixed for each county k with a proportion of $\alpha^{(k)}$, and $1 - \alpha^{(k)}$, respectively. Thus, we identify the two universal patterns that compose the pattern of a county as corresponding to two differently behaving

population groups, whose aggregate tweeting patterns form \mathbf{A} and \mathbf{B} . We have no further restriction on these $\alpha^{(k)}$ values, they can be any arbitrary real numbers.

Then the predicted activity $x_i^{(k)}$ of a county k in hour i would be the following linear combination:

$$x_i^{(k)} = \alpha^{(k)} \cdot A_i + (1 - \alpha^{(k)}) \cdot B_i = \alpha^{(k)}(A_i - B_i) + B_i. \quad (1)$$

Let us denote the weight of each county by $w^{(k)}$, which is proportional to its population $p^{(k)}$, such that $w^{(k)} = p^{(k)} / \sum_{k=1}^M p^{(k)}$. We then define the squared error of our model as

$$E = \sum_{i,k} w^{(k)} \left(y_i^{(k)} - \underbrace{(\alpha^{(k)}(A_i - B_i) + B_i)}_{x_i^{(k)}} \right)^2.$$

We would like to minimize this error with subject to the two conditions $\sum_i A_i = 1$, $\sum_i B_i = 1$. It can be shown (see Additional file 1), that the minimum occurs if $\mathbf{A} - \mathbf{B}$ is parallel to the eigenvector \mathbf{m} corresponding to the biggest eigenvalue of the weighed covariance matrix \mathbf{C} , and that \mathbf{B} can be chosen as the average of $\mathbf{y}^{(k)}$ s. Here, an element of the covariance matrix \mathbf{C} is

$$C_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle, \quad (2)$$

where

$$\langle y_j \rangle = \sum_k w^{(k)} y_j^{(k)}. \quad (3)$$

We now consider a linear representation of the data with a coordinate system where the mean $\langle \mathbf{y} \rangle$ sets the origin and \mathbf{m} is the direction of the line. We calculate $\alpha^{(k)}$ values for each county by projecting $y^{(k)}$ onto this line (see Additional file 1). A positive $\alpha^{(k)}$ means a county, where the majority of people are active on Twitter in correspondence with the daily rhythm dictated by \mathbf{m} , accordingly, negative $\alpha^{(k)}$ is in connection with an opposite pattern.

Because the linear equation system derived from the minimization of the squared error is linearly dependent, the scale on our line is not set (see Additional file 1), as $\mathbf{A} - \mathbf{B}$ is only determined up to an arbitrary scaling factor. Thus, the $\alpha^{(k)}$ values are also determined only up to a scaling factor. Let us now choose \mathbf{A} and \mathbf{B} to be two standard deviations of $\alpha^{(k)}$'s away from the origin $\langle \mathbf{y} \rangle$ in the two directions of our new linear coordinate system:

$$\sigma(\alpha) = \sqrt{\frac{\sum_{k=1}^M (\alpha^{(k)})^2}{M}},$$

$$\mathbf{A} = \langle \mathbf{y} \rangle + 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \quad (4)$$

$$\mathbf{B} = \langle \mathbf{y} \rangle - 2 \cdot \mathbf{m} \cdot \sigma(\alpha). \quad (5)$$

\mathbf{A} and \mathbf{B} are both normalized to 1, where in the 2-dimensional case their components represent the selected two hours, while in the 24 dimensional case they represent the 24 hours of the day. A schematic representation of this model is drawn in Figure 7.

3 Results and discussion

In this section, we present the description and the discussion of the main results of this paper. First, we investigate the correlation between the activities of individual hours and employment and unemployment rates, and choose two dimensions with which employment and unemployment levels have maximum or minimum correlations. We then evaluate to what extent the linear model is a valid description of our data for these most separating dimensions (2) and then for all possible dimensions (24) of our dataset. Second, we discuss how the linear models in 2 and 24 dimensions separate the two population groups with the two distinct activity patterns, and give a possible interpretation of these patterns. Third, we connect the two groups with real-world indicators like share of employed in a county, and discuss the plausibility of the correspondence of the daily patterns of the two separate groups to employment status.

We first evaluate population-weighted Pearson correlations for each hour i between $y_i^{(k)}$ activities for the 1,884 counties (from which we have an adequate number of messages) and employment and unemployment levels. We calculate the errors of these correlations by bootstrapping our sample for $n = 1,000$ times, the results with errorbars are shown in Figure 1. While unemployment levels are defined in the traditional way of the Bureau of Labor Statistics, we define the share of employed slightly differently, normalizing the number of employed by the entire population of a county. This definition matches the notion of population share of “active” people regarding regular working hours better.

The hours between 6am and 8pm show a significantly positive correlation with employment, and a negative one with unemployment, while during the night, between 9pm and 5am, the correlation is reversed. With respect to employment, the correlation peaks at 12pm with 0.43 ± 0.02 and reaches its lowest value at 1am with -0.39 ± 0.03 . The location of the maximum and minimum of correlation with unemployment are shifted slightly to 12pm and 12am, though exactly with opposite signs (0.30 ± 0.02 for 12am and -0.38 ± 0.02 for 12pm). The signs of the correlations and the hours of their extreme values indicate that increased daytime activity is associated with higher employment levels, and higher than average nighttime activity corresponds to higher unemployment.

To check the linearity of the model described in Section 2, we first choose the coordinate system of the hours having the extreme correlation values with employment levels. Figure 2 shows the 12am and 1pm activities of the filtered counties with the dashed line corresponding to the direction of the first eigenvector of the covariance matrix, now calculated only from these two dimensions. If we normalize the eigenvalues by their sum, we see that the first eigenvalue of the covariance matrix carries 0.99 share from all the vari-

Figure 1 Population-weighted Pearson correlation of employment and unemployment levels with hourly activities. Errorbars are calculated using bootstrapping $n = 1,000$ times.

The hours between 6am and 8pm correlate significantly positively with employment and negatively with unemployment. This relationship turns out to be exactly the opposite during the night. Regarding employment, the most distinguishing hours are 1am (most negative correlation) and 12pm (most positive correlation).

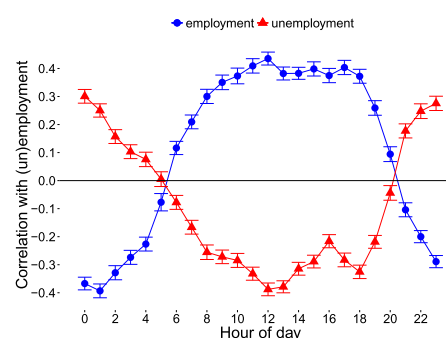


Figure 2 Activity of counties in the space of 12am and 1pm. Each dot represents a county, and the horizontal axis measures the relative tweeting activity between 12am and 1am in that county, while the vertical axis represents the relative tweeting activity between 1pm and 2pm in that county. As these two measures are correlated, a linear transformation could combine them into a single coordinate. The new coordinate axis is represented by the dashed line. The black arrow points to the average of the measures along the original axes. The blue and the red arrows are possible choices for **A** and **B** vectors, see Section 2.4.

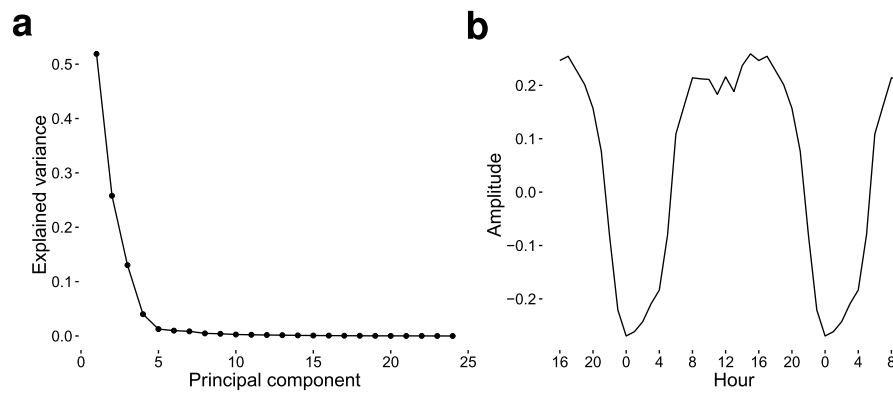
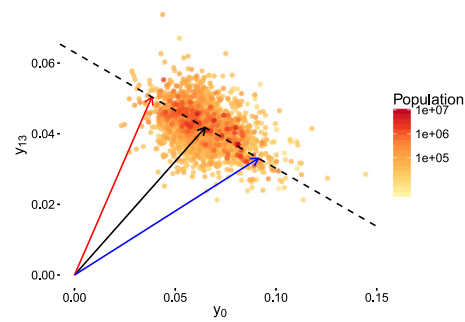


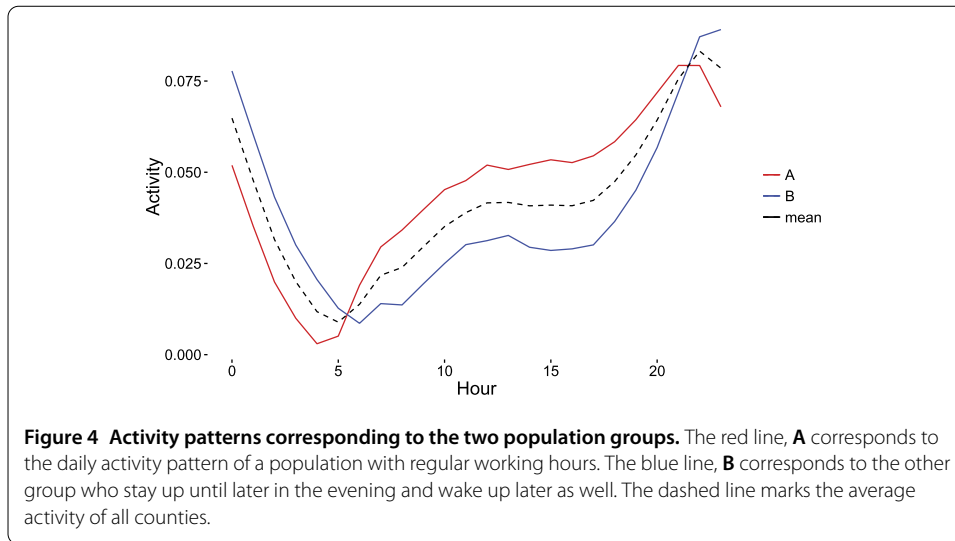
Figure 3 The result of the principal component analysis of the population-weighted covariance matrix. (a) Proportion of explained variance for the principal components of the covariance matrix. Only the first four components carry a share of variance significantly greater than zero. (b) Principal component corresponding to the largest principal value. The amplitude of vector components is plotted, each vector coordinate corresponds to an hour ranging from 0 to 23. The vector components are positive from 5am to 8pm, and negative otherwise.

ance in the data, thus, linearity in this two-dimensional subspace of the whole 24-hour activity space is a good assumption.

We continue by assessing the validity of the linear model in all 24 dimensions presented in Eq. (1). In Figure 3(a) we plot eigenvalues of the covariance matrix C again normalized by the sum of all eigenvalues. Only the first four eigenvalues correspond to a variance significantly greater than 0, and the first principal component stands out with a proportion of 0.52, whereas the other three significant components carry 0.25, 0.13 and 0.04 share of the variance. Thus, our dataset is mostly linear even in the 24-dimensional space, and the representation with Eq. (1) remains plausible.

In the 2-dimensional case, the dashed line of Figure 2 marks the direction of the first principal vector. The difference between the two vectors **A** (red) and **B** (blue) representing the two universal patterns (see Eqs. (4)–(5) in Section 2.4) is parallel to this component, let us denote it by **m**. It can be seen in Figure 2 that the **A** pattern is marked by an increased activity at 1pm, and a decreased activity at 12am, while pattern **B** is characterized by exactly the inverse relationship.

The principal component corresponding to the largest principal value in the 24-dimensional case can be seen in Figure 3. As the coordinates represent the hours, it can be

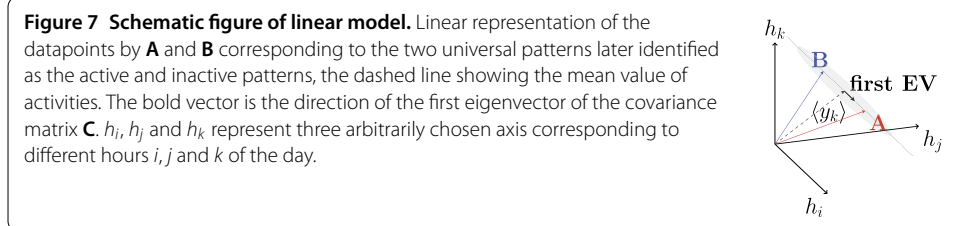
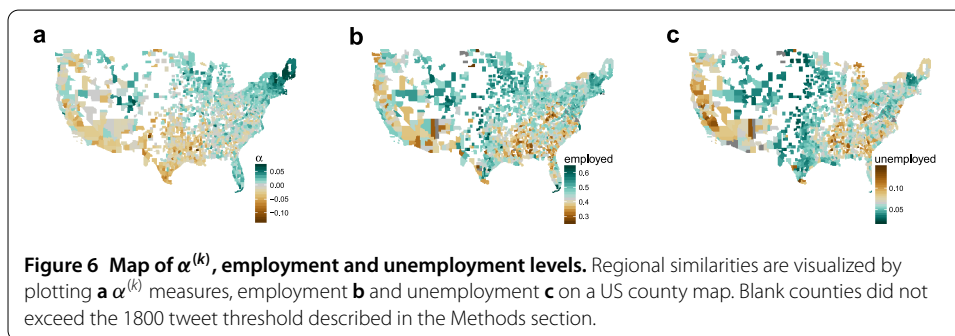
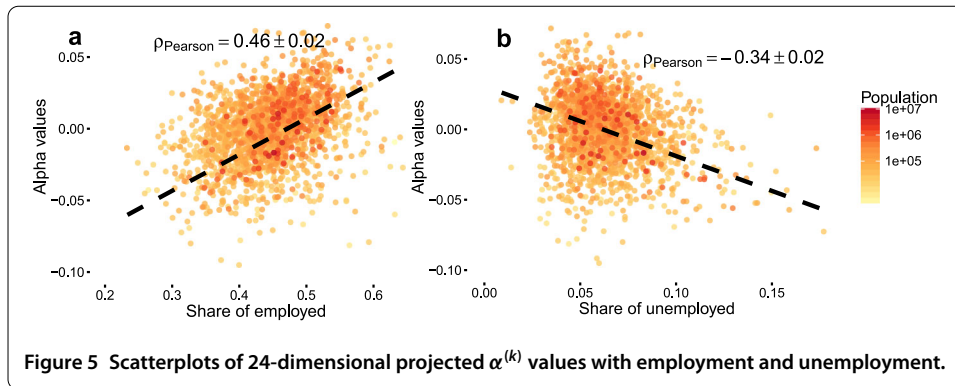


seen from Figure 3 that \mathbf{m} is positive from 5am until 8pm, and negative otherwise. Thus, the positive elements of \mathbf{m} select mainly those hours during which people are awake, and the negative elements correspond to the sleeping hours.

We then plot the elements of the 24-dimensional **A** and **B** from Eqs. (4)–(5) in Figure 4. By interpreting these patterns as the different average tweeting patterns of two population groups, each $\alpha^{(k)}$ is proportional to the share of people in a county in one population group. Our hypothesis is that the group more active during the daytime corresponds to people who regularly go to work, school etc. on weekdays, thus their daytime is regulated by the earlier wake-up and bedtime indicated in pattern **A**. On the other hand, pattern **B** could correspond to a group where this regulation factor does not exist due to retirement, unemployment or any other reason, which would allow these people to be more active during nighttime and wake up later.

To confirm our hypothesis, we correlate $\alpha^{(k)}$ values with labor force and unemployment estimates from the Local Area Unemployment Statistics (see Section 2.2) of the investigated counties. In the 2-dimensional case, these combined values of $\alpha^{(k)}$ do not correlate with employment (0.38 ± 0.03) or unemployment (-0.32 ± 0.02) better than previous activity measures from single dimensions from Figure 1. However, by using all dimensions, we find correlations of 0.46 ± 0.02 and -0.34 ± 0.02 for employment (see scatterplot in Figure 5) and unemployment, respectively. For the employment this is an improvement to that of the single dimensional correlations, while it is not for the unemployment. A possible interpretation is that a stricter daily rhythm is imposed upon those who are employed, as such, the characteristics of their activity curves mean a stronger overall pattern than that of the unemployed. Nevertheless, the result shows that high $\alpha^{(k)}$ is significantly bound to higher employment, and lower unemployment rates, and that the overall shape of the activity timeline can give us more information than just using one feature of a whole day. The similarity of the regional distribution of $\alpha^{(k)}$, unemployment and employment rates are visualized on the three maps of Figure 6.

Our results are in line with previous research carried out for Spain in [31], where share of Twitter activity during a window of the morning hours (8–11am), afternoon hours (3–5pm) and of the night hours (0–3am) correlated significantly with unemployment rates among



25 to 44-year old inhabitants of Spanish administrative areas. High morning and low night activity indicated lower unemployment rates, which is in correspondence with our correlations. Although in Spain high afternoon activity correlated positively with unemployment levels, we cannot observe this phenomenon in the US. Due to the bias in the age of Twitter users towards younger age groups [39], our calculated county activity patterns are not representative of the whole population. We believe that our model could be improved by incorporating labor force data detailed by different age groups.

That correlation with unemployment is significantly lower than correlation with labor force share of the population can be related to the fact that the share of employed should overlap more with the population exhibiting the “working” pattern **A**, whereas officially registered unemployed people are not distinguishable in this context from those who are on a maternal leave or are retired etc. We also believe that there are other inherent reasons for example the more flexible working hours in the creative industry that limit the power of such a simple model explaining the employment patterns of a geographical area.

4 Conclusions

In this paper we analyzed an extensive collection of geolocated tweets originating from the United States between January 2014 and October 2014. We assigned a county to each tweet, then aggregated daily tweeting activity patterns for a typical weekday, and investigated to what extent do hourly activities correlate with employment or unemployment levels. We then modelled daily activity patterns as being the superposition of two universal patterns, thus aiming for a simple linear approximation of our dataset. By minimizing the squared error of our estimations, we obtained that the difference of the two patterns should be parallel to the first eigenvector of the covariance matrix of the dataset and that the mean of the data should fit on our line when selecting only 2 dimensions, and when using all 24 dimensions of our data as well. The set of eigenvalues of the covariance matrix in both cases confirmed the validity of our linear model, which captured most (0.99,0.52) of the variance in the dataset. Whereas in the 2-dimensional case the first eigenvector pointed to the direction, where 1pm activity was increased, and 12am activity decreased, in the 24-dimensional case it had positive elements during the daytime hours (6am-8pm), and was negative during the most of the night (9pm-5am).

By projecting county activity patterns onto these lines with the mean as the origin, we obtained a measure for each country that indicated the extent to which the tweeting pattern of a county resembles that of the first eigenvector. This measure has been shown to correlate significantly with county labor force shares and unemployment rates, though in the 2-dimension case, these correlations could not enhance the performance of the single hourly correlations. Using all 24 dimensions, we obtained a better Pearson correlation of 0.46 ± 0.02 and -0.34 ± 0.02 for employment and unemployment, respectively. The signs of the correlations indicate a relationship where counties exhibiting a higher tweeting activity during the daytime (6am-8pm) have higher employment and lower unemployment rates, and counties with increased night activity can be related to lower employment and higher unemployment rates. These correlations show, that even though Twitter population is biased towards younger age groups, and employment data was considered for all age groups, the underlying relationship between daily activity patterns and employment data can be captured with plausible outcomes.

Our results thus showed, that by analyzing a relatively sparse publicly available geolocated dataset, a very simple model can explain to a significant extent such an important socio-economic indicator as employment/unemployment. We believe that our model can be even further improved by incorporating detailed data for different age groups or other datasets from either traditional or digital sources such as mobile traffic data. It would be worth to investigate whether dynamic changes of activity patterns over time can follow employment trends. This kind of analysis would allow policy makers a better insight into the processes connected to employment phenomena, and could form the basis of future datasets, where problems could not only be identified based on officially registered unemployed people, but also on a basis of the digital footprints people leave on different platforms.

Additional material

[Additional file 1: Technical details of the linear model calculations in this article \(pdf\)](#)

Funding

The authors received partial support of the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2-C-11/1/KONV-2012-0013), Ericsson, the OTKA-114560 and the MAKOG Foundation.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the following repository:
<http://www.vo.elte.hu/papers/2016/unemployment/>.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GV conceived the experiment, EB and ZL collected the data, EB and GV analyzed the results, EB wrote the manuscript. All authors reviewed the manuscript.

Publisher's Note

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

Received: 23 November 2016 Accepted: 28 June 2017 Published online: 11 July 2017

References

- Aschoff J, Wever R (1976) Human circadian rhythms: a multioscillatory system. *Fed Proc* 35(12):2326-2332
- Cagnacci A, Elliott JA, Yen SS (1992) Melatonin: a major regulator of the circadian rhythm of core temperature in humans. *J Clin Endocrinol Metab* 75(2):447-452. doi:10.1210/jcem.75.2.1639946. PMID:1639946
- Refinetti R, Menaker M (1992) The circadian rhythm of body temperature. *Physiol Behav* 51(3):613-637. doi:10.1016/0031-9384(92)90188-8
- Cajochen C, Kräuchi K, Wirz-Justice A (2003) Role of melatonin in the regulation of human circadian rhythms and sleep. *J Neuroendocrinol* 15(4):432-437. doi:10.1046/j.1365-2826.2003.00989.x
- Taillard J, Philip P, Bioulac B (1999) Morningness/eveningness and the need for sleep. *J Sleep Res* 8(4):291-295. doi:10.1046/j.1365-2869.1999.00176.x
- Aledavood T, Lehmann S, Saramäki J (2015) Digital daily cycles of individuals. *Front Phys* 3(October):Article ID 73. doi:10.3389/fphy.2015.00073
- Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *Eur Phys J B* 88(6):Article ID 164. doi:10.1140/epjb/e2015-60106-6
- Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census. *IEEE Pervasive Comput* 6(3):30-38. doi:10.1109/MPRV.2007.53
- Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the space - time structure of the mobile phone network. *Environ Plan B, Plan Des* 36(5):824-836. doi:10.1068/b34133t
- Calabrese F, Reads J, Ratti C (2010) Eigenplaces: segmenting space through digital signatures. *IEEE Pervasive Comput* 9(1):78-84. doi:10.1109/MPRV.2009.62
- Soto V, Frías-Martínez E (2011) Automated land use identification using cell-phone records. In: *Proceedings of the 3rd ACM international workshop on MobiArch - HotPlanet '11*, pp 17-22. doi:10.1145/2000172.2000179
- Toole JL, Ulm M, González MC, Bauer D (2012) Inferring land use from mobile phone activity. In: *Proceedings of the ACM SIGKDD international workshop on urban computing - UrbComp '12*, pp 1-8. doi:10.1145/2346496.2346498
- Pei T, Sobolevsky S, Ratti C, Shaw S-L, Li T, Zhou C (2014) A new insight into land use classification based on aggregated mobile phone data. *Int J Geogr Inf Sci* 28(9):1988-2007. doi:10.1080/13658816.2014.913794
- Louail T, Lenormand M, Picornell M, García Cantú O, Herranz R, Frías-Martínez E, Ramasco JJ, Barthélemy M (2015) Uncovering the spatial structure of mobility networks. *Nat Commun* 6:Article ID 6007. doi:10.1038/ncomms7007
- Grauwin S, Sobolevsky S, Moritz S, Gódor I, Ratti C (2015) Towards a comparative science of cities: using mobile traffic records in New York, London, and Hong Kong. In: *Computational approaches for urban environments*. Springer, Cham, pp 363-387. doi:10.1007/978-3-319-11469-9_15
- Kondor D, Thebault P, Grauwin S, Gódor I, Moritz S, Sobolevsky S, Ratti C (2015) A tale of many cities - visualizing signatures of human activity in cities across the globe. *Landsc Archit Front* 3(3):54-61. doi:10.1007/slaf-0054-15008
- Lenormand M, Picornell M, Cantú-Ros OG, Louail T, Herranz R, Barthélemy M, Frías-Martínez E, San Miguel M, Ramasco JJ (2015) Comparing and modelling land use organization in cities. *R Soc Open Sci* 2(12):Article ID 150449. doi:10.1098/rsos.150449
- Cici B, Gjoka M, Markopoulou A, Butts CT (2015) On the decomposition of cell phone activity patterns and their connection with urban ecology. In: Shen S, Sun Y, Chen J, Zhang J, Zussman G (eds) *Proceedings of ACM MobiHoc '15*. ACM, New York, pp 317-326. doi:10.1145/2746285.2746292
- Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási A-L (2008) Uncovering individual and collective human dynamics from mobile phone records. *J Phys A, Math Theor* 41:Article ID 224015. doi:10.1088/1751-8113/41/22/224015
- Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2014) High resolution population estimates from telecommunications data. *EPJ Data Sci* 4(1):Article ID 4. doi:10.1140/epjds/s13688-015-0040-6
- Schneider CM, Belik V, Couronne T, Smoreda Z, Gonzalez MC (2013) Unravelling daily human mobility motifs. *J R Soc Interface* 10(84):Article ID 20130246. doi:10.1098/rsif.2013.0246
- Smith-Clarke C, Mashhadi A, Capra L (2014) Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp 511-520. doi:10.1145/2556288.2557358
- Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, Pentland A (2014) Once upon a crime: towards crime prediction from demographics and mobile data. In: *Proceedings of the 16th international conference on multimodal interaction - ICMII '14*, pp 427-434. doi:10.1145/2663204.2663254

24. Jiang S, Ferreira J, González MC (2012) Clustering daily patterns of human activities in the city. *Data Min Knowl Discov* 25(3):478–510. doi:10.1007/s10618-012-0264-z
25. Ettredge BM, Gerdes J, Karuga G (2005) Using web-based search data to predict macroeconomic statistics. *Commun ACM* 48(11):87–92. doi:10.1145/1096000.1096010
26. Choi H, Varian H (2012) Predicting the present with Google trends. *Econ Rec* 88(Supplement s1):2–9. doi:10.1111/j.1475-4932.2012.00809.x
27. Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: evidence from the Visegrad Group countries. *PLoS ONE* 10(5):Article ID e0127084. doi:10.1371/journal.pone.0127084
28. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. *Pers Ubiquitous Comput* 10(4):255–268. doi:10.1007/s00779-005-0046-3
29. Proserpio D, Counts S, Jain A (2016) The psychology of job loss: using social media data to characterize and predict unemployment. In: Nejdil W, Hall W, Parigi P, Staab S (eds) *WebSci*. ACM, New York, pp 223–232
30. Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D (2015) Tracking employment shocks using mobile phone data. *J R Soc Interface* 12(107):Article ID 20150185. doi:10.1098/rsif.2015.0185
31. Llorente A, García-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. *PLoS ONE* 10(5):Article ID e0128692. doi:10.1371/journal.pone.0128692
32. Morstatter F, Pfeffer J, Liu H, Carley K (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In: *International conference on weblogs and social media*, pp 400–408
33. Dobos L, Szule J, Bodnar T, Hanyecz T, Sebok T, Kondor D, Kallus Z, Steger J, Csabai I, Vattay G (2013) A multi-terabyte relational database for geo-tagged social network data. In: *Proceedings of the 4th IEEE international conference on cognitive infocommunications - CogInfoCom 2013*, pp 289–294. doi:10.1109/CogInfoCom.2013.6719259
34. Szalay AS, Gray J, Fekete G, Kunszt PZ (2007) Indexing the sphere with the hierarchical triangular mesh. *arXiv:cs/0701164*
35. Kondor D, Dobos L, Csabai I, Bodor A, Vattay G, Budavári T, Szalay AS (2014) Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh. In: *Proceedings of the 26th international conference on scientific and statistical database management - SSDBM '14*, pp 1–4. doi:10.1145/2618243.2618245
36. Global Administrative Areas. <http://gadm.org>. Accessed 21 Nov 2016
37. United States Census 2010. http://www2.census.gov/census_2010/. Accessed 21 Nov 2016
38. Local Area Unemployment Statistics. <http://www.bls.gov/lau/>. Accessed 21 Nov 2016
39. Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M, Rainie L, Smith A (2015) *Pew Social Media Report 2015*. Technical report, Pew Research Center

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com