Trends in urban flows: from Wi-Fi data to pedestrians' route choices

I Introduction. The accurate estimation of human activity in cities is one of the first steps towards understanding the structure of the urban environment. Human activities are highly granular and dynamic in both the spatial and temporal dimensions and estimating confidence is crucial for decision-making in numerous applications such as urban management, retail, transport planning and emergency management. Detecting general trends in the flow of people between spatial locations is neither obvious nor an easy task due to the high cost of capturing these movements without compromising the privacy of those involved. This research intends to address this problem by examining the movement of people in a SmartStreetSensors network at a fine spatial and temporal resolution. A novel methodology to the field of Big Data using mathematical models from information theory is introduced.

II Data. The SmartStreetSensor project is one of the most comprehensive studies undertaken on footfall (FF) volume and characteristics in urban areas across the United Kingdom. The data for the study is generated through sensors installed at around 1000 locations across the UK. These sensors capture a series of public signals - known as probe request frames - generated by Wi-Fi enabled devices. In particular, the number of probe requests captured during 2017 was ~53 million records corresponding to all the aggregated 5-minutes FF counts at 889 locations (35 % of them in the Greater London area). Each location generates a daily FF signal like the one shown in Figure 1. The initial visual inspection and analysis of the footfall signatures demonstrated that Mondays through Thursday display different patterns to those observed during the weekends and, moreover, even though Fridays are mostly similar to the rest of the workdays, some locations may have a pronounced activity in the evening hours.

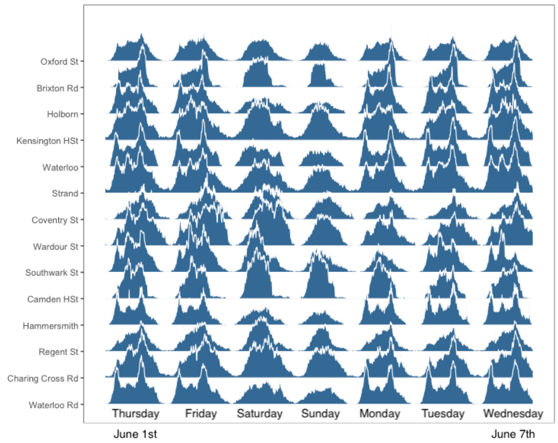


Figure 1. Different locations in London generate different FF signals. The difference between weekdays and weekends is clear. The shape on weekends tends to be more like a plateau while weekdays exhibit one, two or three clear peaks.

III Results. A) Areas Classification. Although Fig. 1 provides some clues about the different activities in different locations, this is only a visual analysis. To formalise this idea, we produce the typical weekday / weekend signal for each one of the 889 locations and then cluster all the resulting time series following a Dynamic Time Warping (DTW) technique, finding a core of eight canonical FF signals that tell us something about the activities around different locations (Figure 2).

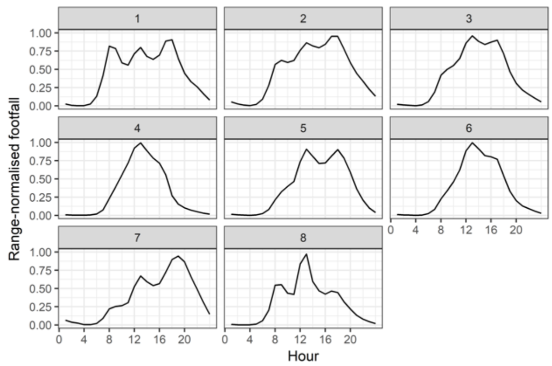


Figure 2. Classification of Weekdays FF signals

For example, profile 1 (Commute and lunch) is three-peaked profiles characterised by busier FF traffic at all three characteristic periods during the day - morning rush hour, lunchtime and afternoon rush hour.

B) Pedestrians Flows. Until this point, we only have analysed isolated locations, but is clear that in the complex urban landscape, nearby locations have some degree of influence onto each other, and that in the FF counts, a percentage of the pedestrians counted at location S1 at time t would be the same counted at location S2 at time t+n. The aim is to, without tracking people provide a measure for the size of the flow between each pair of sensors.

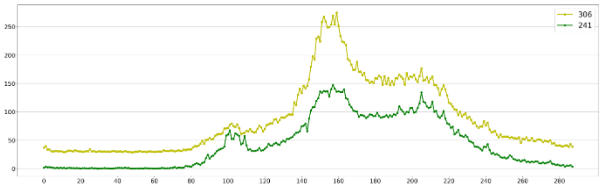
One way to accomplish this is to think of this motion of people as flows of information among distinctive sources, so we can relate the number of people reaching one sensor from another by measuring the uncertainty between two interacting random variables. For this, we used an information theory concept known as Local Transfer Entropy LTE defined by

$$LTE(t+1, x_{j-1}, y_j) = \log_b\left(\frac{p(y_{t+1} \mid y_t, x_t)}{p(y_{t+1} \mid y_t)}\right) \qquad (1)$$

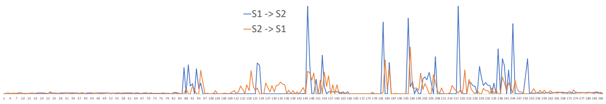
Where t indicates a given 5-min bin, and x, y the corresponding values at S1, S2 at t. Eq. 1 measures the reduction in uncertainty at yt, given xt and yt+1 in comparison with the case when the only xt is known. In general, the LTE would inform not only if there is a strong correlation between two signals, but the direction in which the information is flowing, at every time bin.

For example, Fig. 3.a shows a pair of sensors with a correlation=0.76 and 3.b the LTE between both signals, with the blue line representing the flow of information from S1 to S2 and the red one the flow from S2 to S1. After carefully comparing the LTE values at each bin, we produce a metric that tells us the overall behaviour of these flows for all pairs. With this, we could explain the high/low correlation and estimate the direction of the pedestrian flows for around 70% for the pairs.

C) Route complexity. Finally, we notice that for 40% of the pairs, the LTE profiles were the same between two pairs or meaningless as the LTE values were the same or 0.



a)



b)

Figure 3. a) FF signals at two locations at 3 min. walking distance from each other; b) LTE for the same locations.

So, it must be something else. We decide to perform a semantic analysis over the walking directions between pairs of locations, obtained from the Google API. This service response with a series of instructions that share a set of five basic commands: **Walk, Continue, take (1), Slight (2), Turn (3), and Take, follow by the word bridge of zebra (4)**. We then isolated these words from each set of instructions and assigned a weight to each one of those five words:

$$Score_r = \sum_{i=1\ldots4} W_{ir}/t \qquad (2)$$

Where Wir is the value in parenthesis after each word, Eq.2 is just the sum of these values, weighted by the time that takes to complete each one. In this way, routes more complex to navigate (with a lot of turns or crosses) will have a higher score that straight line routes.

Conclusions

1. We can use probe requests as a proxy to detect human activity.
2. We can use FF signals to classify areas.
3. The relation between two signals is NOT in function of the walking.
4. We can have a feeling about the flow of people between two locations using LTE.
5. Routes complexity could probably explain the correlations, but doesn't tell us anything about the flows