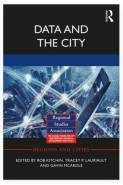

Data and the City
Edited by Rob Kitchin, Tracey P. Lauriault, Gavin McArdle
232 pages | 29 B/W Illus.
e-Inspection Copy
Routledge
Taylor & Francis Group

3 Data about cities

Redefining big, recasting small

Michael Batty

Introduction

Prior to the industrial revolution, record-keeping was an intensive but modest affair with manual technologies constraining the growth of data. The development of mechanical technologies from the late eighteenth century began to change this and local records gradually became more automated during the nineteenth century. The Population Census was in fact one of the only systematic catalogues of data produced on a continuing basis at a national level until national accounts and related economic data began to be collected seriously and routinely in the 1920s (Bos 2011). Automation, however, using mechanical devices continued apace in the early twentieth century and the first digital computers in mid-century embraced the challenge of dealing with ever larger data volumes that now form the basis of all kinds of development in electronic media and communications technologies.

Historically, data were always big with respect to the available means by which they could be manipulated. There is a wonderful story from the 1950s about the use of spare cycles in the early computers developed for the Lyons Tea Company (Ferry 2010) where these computers were used to compute shortest routes for freight in the rail system so that British Railways could price these goods accordingly. Dramatic and ingenious manipulations had to be devised to make this possible, such as stuffing data and intermediate calculations into all corners of memory and Scotland needing to be treated separately from the rest of Britain and then stitched back together after separate computation. In the process, those involved actually invented the well-known Dijkstra algorithm a year before Dijkstra did so himself and some four years before he published it (Graham-Cumming 2012). Countless examples such as these exist, which show how the limits of computation were reached with new algorithms, and data mining techniques were invented on the back of data which were then viewed as ‘big’.

So ‘big’ with respect to data is a relative concept and some data have always been big with regards to how they might be manipulated using state-of-the-art computation. But apart from the sheer volume of data, in cities data have always been big in another sense. Here, our concern is no longer with location but with interactions (Batty 2013): relationships between locations are best expressed by flows. The volume of data contained in flows is, in general, the square of

the elements that define the locations between which the flows are generated. If there are n locations, then there are n^2 possible interactions between them and thus the data associated with interactions increases exponentially as the number of locations increases or as locations get finer and finer in terms of their resolution. Here, the contention is that big data can be generated from small data through interactions, and that higher order effects are in fact big data. Although I do not conclude that the big data revolution is a red herring, we will conclude that 'bigness' is never what it seems and that 'bigness' in terms of computational time taken to explore data, which might be quite small in size, is as important as dealing with massive data volumes.

Classifying city data: the data cube

Introduced by Brian Berry (1964), an early data typology that has withstood the test of time is the 'geographic matrix'. This consisted of an array of places – locations – and their attributes, which he called characteristics. Such a matrix, he argued, was the essence of geographical analysis in that the dimension of place and its characteristics or attributes defined the central qualities of any location. To this he added another dimension, time, though this rarely had the same level of detail of the other two. In fact, he envisaged these additional time slices to be limited in number, though in principle each of these dimensions could take on any number of categories. Although he did not use the term, the geographic matrix in its three-dimensional form is close, if not identical to, what in data science is now called the 'data cube' (Han *et al.* 2011). Berry then proceeded to use this matrix to explode a spatial data set. In one sense, the focus was on place rather than its characteristics or its temporal positioning, but by concatenating these dimensions one might envisage a series of relationships in single, pairwise or in three-wise fashion. If we label characteristics by their volume as M, places as N, and time slices by T, then there are seven possible combinations of relations: M, N and T by themselves, $M \otimes N$, $M \otimes T$ and $N \otimes T$, and $M \otimes N \otimes T$. Unpacking these further, we might consider relations between $M \otimes M$, $N \otimes N$, and $T \otimes T$. Significant for this discussion is the relation between N and itself which essentially is spatial interaction – linkages or flows between locations. Berry's focus however was on another kind of data explosion that comes from generating relationships between the dimensions. We will illustrate these here with respect to relationships between places – spatial interactions – which can also be tagged to quite fine resolutions of time.

In fact, it is important to be clear as to the way the data cube might be used in the analysis of city data. Even though it is based on three dimensions, which can in fact be extended to many more, usually any analysis takes one of these as being the anchor point – place, characteristics or time – and conducts analysis with respect to relationships associated with this anchor. Although the data cube is generic, whenever data are considered in these terms, the problem is usually structured from one of these perspectives and thus it is important to see the size of data, its volume and its variety at least in terms of the particular perspective adopted. It is worth indicating how traditional urban data – urban populations collected from traditional sources such as complete Population Censuses – can explode

into big data. This was possible long before the current era and it is very obvious when spatial interaction is considered. In 1964, Lowry built a state-of-the-art urban model for Pittsburgh which divided the region into 456 zones between which the flows of people moving to work, shop and so on were collected. The data were collected from household interviews intended for traffic studies, but the volume when considered with respect to the matrix of interactions was huge by the standards of those times where $456^2 = 207,936$ possible interactions (trips) was standard. This was in an era when many mainframe computers could barely store more than 64K numbers and most of the transport models then built always pushed up against these limits. Indeed, it was one of the main reasons for the enormous problems associated with the earliest urban models, which Lee (1973) in his famous paper defined as one of data 'hungriness' (Batty 2014).

The emergence of big data in cities

Before turning to examples, it is important to get a tangible sense of what the term 'big data' means, for it has only become significant in the last decade. This has coincided with the development and dissemination of countless digital devices that sense characteristics of objects in the physical environment with respect to their type, positioning and the time when they are observed. These are, of course, the three dimensions of our data cube and big data thus tends to be data that are dimensioned in at least these three ways – by their attributes or characteristics, by their spatial positioning or location, and by the time instant at which the relevant objects are observed. The objects can be human or physical, indeed of any type as long as they are associated with a relevant sensing device.

There are many definitions of big data. The cliché is that big data are defined by volume, variety, velocity, veracity and value. This simply roots the data in questions of size (bigness), variety (diversity and extent), velocity (temporal frequency of collection or observation), veracity (level of accuracy and/or uncertainty) and value (what it brings to various purposes), but it might be objected that all these criteria apply to small data. However, the implication is that it is size, scale and scope that pertain to these characteristics (IBM n.d.). In fact, big data are much more than these four or five 'Vs'. Dutcher (2014) has collected together some 40 definitions from 'thought leaders' across the industry and one of the main conclusions is that big data are more about the tools that are needed to process them than their size or volume.

Often big data are hard to understand because they have little structure, they are sometimes but not always large, and traditional tools are very difficult to use in their processing. For example, very large quantities of household census data, although not any larger in the volumetric sense than at any time in the last half century, often stretch and confuse traditional multivariate techniques. Even plotting a scatter diagram relating, say, population income to level of education at the individual or household level for a country the size of the UK requires visualizations of more than 20 million points and most if not all statistical packages and even statistical interpretations break down when confronted with such data volumes. Even so, such data would not be regarded as big data by contemporary

standards for the usual rule of thumb is that the data must be giga- and upwards in size for it to be classed as big data.

Big data which are streamed in real time represents the cutting edge of new data about the functioning of cities. Much of these data are streamed from devices that are simply embedded in the physical environment and transmit data in continuous fashion with little human interference or management, such as loop counters which record traffic volumes, digital weather stations, and such like. Much of these are captured in the various dashboards that have been set up to pull together such data and make them intelligible to interested observers and policymakers. These dashboards have mainly been produced so far to demonstrate that by visually synthesizing such data one can gain an immediate impression of the state of the city (O'Brien *et al.* 2014; Kitchin *et al.* 2014). In fact, the synthesis that is required to make sense of this is very hard to develop as many of the data sources cannot be easily integrated. Moreover, much of these streamed, real-time data reflect very different concerns for cities from more traditional data sets.

Real-time data pertaining to the socio-economic structure of the city are much more problematic to collect using sensing devices. Unambiguous answers to queries which involve the human condition are almost impossible to link to real-time sensors. Information on people's choices are fraught with difficulty in terms of collection and interpretation. The reason why so much data in real time are transit data is that travel is a relatively routinized activity, whereas collecting data about unemployment, income, employment activity, migration and so on requires human and related agencies to put in place systems where people are required to respond by answering or registering. Some data are being picked up in retailing with respect to sales data from smart, credit, loyalty cards and so on, but invariably where these data are collected (and sometimes available) in real-time, various sensing devices are used. Data which are compiled from registrations are increasingly being made in near real-time, such as house prices. In these cases, the frequency at which such data are produced is monthly, possibly weekly at best to date, but these kinds of data depend on the frequency of changes – people make changes in these phenomena over matters of days and weeks and months rather than seconds and minutes (Batty *et al.* 2015).

To illustrate these issues, we will focus on transport where data are intrinsically big, including traditional data collected from questionnaires about travel patterns administered to individual travellers or households, smart card usage for collecting fares, real-time movement data from vehicles themselves, and data captured by monitoring passengers using automated observations. Not only are transport data big in that much of them deal with how travellers move between origins and destinations, thus generating spatial interactions, but they are also big temporally because automated methods can capture data continuously.

Traditional transport interaction data: big data generating complex visualizations

Ever since transportation planning formally began in the 1950s, the focus has been on potential interactions or flows between origins and destinations. Different types

of traffic form the essence of transport models, usually based on different modes, but the class of models that we will allude to deal with many other kinds of flow from social networks, to input–output trade relations, to patterns of migration, and so on. The concatenations that we are focusing on here are flows between places, that is $N \otimes N$ which generate travel volumes that can be substantial as the number of places N increases, as we noted above for Lowry's (1964) model of Pittsburgh. Until quite recently visualizing flows has been stymied by constraints imposed on graphics. To consider the nature of the problem, in Figure 3.1(a) we show London divided into 33 separate but contiguous zones for which a journey to work matrix – flows from any zone (which is an administrative borough) to any other – is almost impossible to plot clearly. Thirty-three zones generate a total possible number of trips $33^2 = 1089$ which may not appear to be a large number, but is hard to plot clearly. We show this plot in Figure 3.1(b) where plotting all links from any zone to another, but excluding the intra-zonal trips and also suppressing the asymmetry of the matrix where the flow from zone i to j is generated by adding the flows as $T_{ij} + T_{ji}$, still produces a map which is hard to interpret. Plotting individual trips from one origin to all destinations is the only way to make the map clear but we get no sense of the polycentricity of the system from this visualization and this is what we really need to detect in the data.

Now this is a very crude characterization of the journey to work in Greater London. Even 50 years ago, we would not be content with this level of resolution and therefore we will need to work with a much bigger data set by dividing these 33 zones into their constituent wards – typically local electoral districts which have on average around 13,500 residents living within them. There are 633 such zones and immediately the data have exploded to $633^2 = 400689$ potential interactions, which is quite large. We usually calibrate a model for this kind of data so that we predict each of these flows, but many of the flows for a system of this size and resolution will be small and quite a few zero in terms of the observations.

In Figure 3.2, we show the more disaggregate zoning system. It is not worth showing a plot for the full trip matrix as this is simply a mess with no way of detecting the complexity of the physical form. What we want to do is detect how close different patterns from different parts of the metropolis are and a first way into this problem is visualization. The notion of examining trips origin by origin or destination by destination is an obvious way forward and we do this in

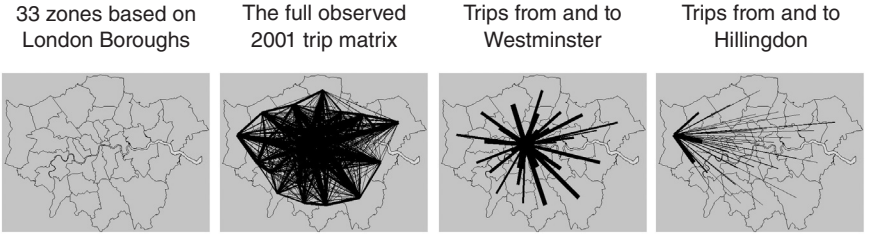

Figure 3.1 Total two-way trips: a) the zoning system; b) all trips plotted; c) trips associated with Westminster (the centre); d) trips associated with Hillingdon (Heathrow). Note that intra-zonal trips are not plotted

Figures 3.2(b) and (c) as we did in Figure 3.1 for the coarser resolution system. Aggregation and animation are ways of dealing with these data in terms of building up a structured understanding of this complexity, but the problem really becomes serious once we wish to test comparisons and compute correlations between the observed trip matrix and any other matrix such as a predicted one. To show how this kind of problem explodes into big data, which need new methods, we will compare the 633 × 633 matrix with one that is predicted by the model.

We now need to note the model that we will build to produce the predictions to be compared against the data in Figure 3.2. The model predicts trips T'_{ij} between origins O_i^{obs} and destinations D_j^{obs} which are then compared against observed trips T_{ij}^{obs}. Observed origin and destination volumes O_i^{obs} and D_j^{obs} are computed from the observed data as $O_i^{obs} = \sum_j T_{ij}^{obs}$ and $D_j^{obs} = \sum_i T_{ij}^{obs}$. The model is an unconstrained gravity model that computes predicted trips as a function of the observed origin and destination volumes and an inverse function of distance d_{ij} between each origin and destination pair. The model can be stated as $T'_{ij} = K\, O_i^{obs} D_j^{obs} \exp(-\beta d_{ij})$ where K and β are parameters that meet normalizing constraints. From the model, we clearly derive predicted trips but also predicted origin and destination totals $O'_i = \sum_j T'_{ij}$ and $D'_j = \sum_i T'_{ij}$. To measure the goodness of fit of the model with the data, we need to examine the scatter plots which contain the correlations between O'_i and O_i^{obs}, D'_j and D_j^{obs}, and T'_{ij} and T_{ij}^{obs}.

The scatter plots for origins and destinations are easy enough to visualize as there are 633 observations in each. However, for the trips, there are a possible total of 400,869. In terms of the observed trip data, some 64 per cent of these are zero, and as the data are taken from a 10 per cent sample, this poses a problem. Should we compare zero cells with predicted ones, which will always be positive, and should we compare cells with a fractional number with integers? If we exclude the zero cells, then we still have some 142,291 to deal with, implying that only 36 per cent of our data matrix is occupied. We illustrate these patterns in Figures 3.3 and 3.4.

Figure 3.3 is revealing. The three scatters are very different with employment being predicted rather well, residential population less well, and trips showing

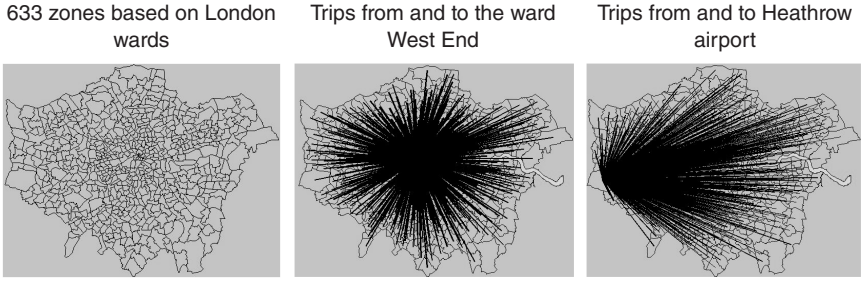

Figure 3.2 Total two-way trips: a) the fine-scale zoning system, b) trips associated with an inner-city ward, c) trips associated with Heathrow airport

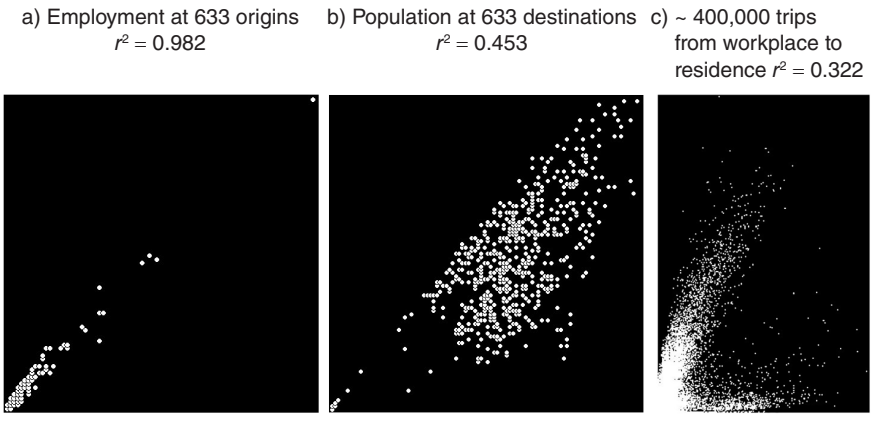

Figure 3.3 Predicted against observed data: a) origin employments; b) destination working populations; and c) trips from work to home

that there are at least two regimes characterizing travel in London. In fact, the scatter of trips in Figure 3.3 reveals a clear density map and in Figure 3.4 we show this as best we can. The intensity of very small trips is much greater than larger ones for the distribution of trip volumes follows some sort of power law.

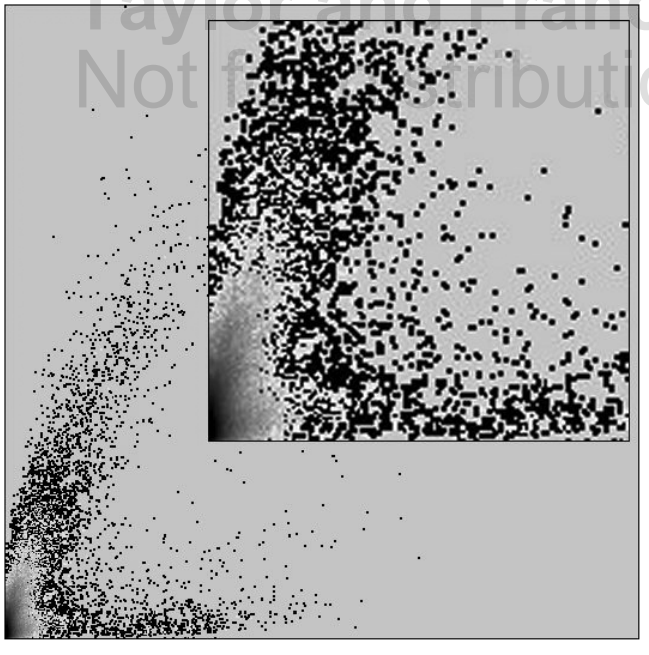

Figure 3.4 The density of the scatter: different patterns at different scales

In Figure 3.4, we have blown up the lower portion of the scatter to reveal this intensity and this reveals that this kind of data mining must be supplemented by many other kinds of visualization and analysis so that the true patterning of a system with this kind of complexity can be laid bare.

Now all this may not look very much like big data, but our current extensions of these models are equivalent to entire systems of cities at the same level of resolution as the Greater London model zoning system in Figure 3.2. We are now working on a model with 7,201 zones which have an average population for England and Wales of some 7,000. Our model is built for all these zones and immediately there comes a problem of visualizing the scatter of origins and destinations as well as trips of which there are a total possible cells in the matrix of $7,201^2 = 51,854,401$. Visualizing nearly 52 million points on a scatter graph is well beyond our capabilities and although only 10 million or so of these points are likely to be above zero, this is still beyond the capabilities of this kind of analysis. We show the zoning system in Figure 3.5(a) and when we move to flows, it is impossible to use the single origin, many destination tool to visualize a set of flows one by one. What we have done here is to produce a single flow for each origin to all its destinations using a weighted directional vector. For each origin i, we compute the average vector as a single arrow showing the average strength and direction as $[\vec{x}_i, \vec{y}_i] = [(x_i y_i), (\Sigma_j T_{ij}[x_i - x_j]/n, \Sigma_j T_{ij}[y_i - y_j]/n)]$. Much information is lost in our visualization but in the system we are developing, there is zoom capability that is able to illustrate the overall pattern at a coarse spatial scale and the detail at the finest scale of the zones themselves. We show the coarser visualization for England and Wales in Figure 3.5(b).

The zoning system for England and Wales

Average directional flows from population centres to employment in E&W

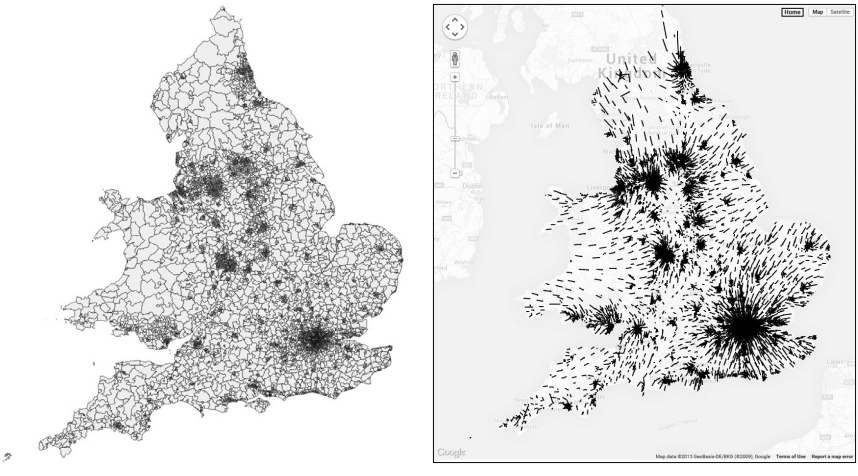

Figure 3.5 Visualizing big data in tens of millions or more of transport flows

Much of this has been possible in terms of data available for the last 30 years or more but only now that we have computers large enough are we able to exploit the bigness of these data. This is very different from the big data that we will present in the next section where the volume comes largely from the temporal and individual rather than spatial dimension. It does reveal, however, that big data have been with us for a while and it is computation more than anything else that determines the size of data set that we can handle, interpret and use fruitfully.

Real-time streamed transportation data at the micro-level

Since the 1950s, data have been collected in continuous time for traffic flow analysis. Much of these data have been hard to link to origin-destination data of the kind just examined largely because they are supply-side data pertaining to vehicular movement and not to intentional trip-making. However, with the advent of RIFD and related technologies, it is now possible to collect data on where people enter and exit a transit system or where they embark and end any journey if the relevant collector is in place. Devices which are specially devised for the data collection in question are by far the best as the data that they produce are unambiguous (although there may be substantial noise still to be filtered out). Mobile devices for other purposes, such as phones, can also be used to extract data from call detail records which locate the phone when a call is made (Chen *et al*. 2015).

Because these data are recorded at the exact time when the smart card or mobile device is linked to the system in question, there is a continuous or at least continual record of activations which represent real-time collection, either accessible in real-time itself or for *post hoc* analysis. In short, the data are as voluminous as the number of activations. If this is phone calls, then it is the number of calls made from that device per day or over whatever unit of time and space the data are aggregated to. Here, we will use data generated by the Oyster card, a RFID smart card used on all public transport in Greater London. This card stores the money that travellers use to pay for journeys and the system is designed to recognize the category of payer as well as the time and place where the traveller taps in or out of the system. Travellers tap in and out on trains but only tap in on buses.

We have several tranches of data from this system. Our largest set is for 86 days in the summer of 2012 where there were 9,902,266,857 (nearly 10 billion) taps. Of these taps, 44 per cent were on buses and 56 per cent on rail, which is tube and overground with some being on the mainline network rail. As there is only tap-in on buses, we can guess that if round trips are made by rail, then this is about half of all rail trips meaning that there are about 60 per cent more bus trips than rail. The data also show that 11,535,090 different Oyster cards are used for these 10 billion taps, which is 86 taps per unique card, on average about one per card per day.

These data are quite unstructured. They come as a flat files where each tap is recorded by place and time – subway station, location of bus by stop, etc., and some classification of the traveller such as whether the card is free, and what the payment category is. Generally, it is possible to trace the behaviours of an individual cardholder through time and space. The degree of heterogeneity in the data set is

enormous and this is a feature that makes them usable for all kinds of temporal modelling at the level of the cardholder conceived of as an agent. However, there are critical problems. The analysis of one day's worth of data in November 2010 from a series we have of three weeks' data for the 660 tube and overground rail stations revealed that 6.2 million travellers tapped in but only 5.4m tapped out. Essentially this was because barriers were up. A large class of Oyster users with free passes are not fined for not tapping in or out while season ticketholders are also not fined as their cards are loaded with a fixed amount of money for a period. This is quite a large loss of data. If you combine this with travellers using more than one card, then this confounds the data set for transport analysis.

It is possible with some analysis to figure out how many journeys are made by tracing different travellers in terms of the tap-in and -out activity during the working day, for rail at least. We have attempted some analysis of buses with respect to travellers who have a unique identifier and who hop onto buses and trains within a certain time interval, which we assume captures some multi-modal journeys, but our analysis is limited and our confidence in extracting multimodal journeys is low. In terms of the rail system, we are able to produce distinct trips in terms of segments although the analysis of round trips is more limited. For example, in the 2012 data, we can identify 291 million trips between one station and another in terms of a tap-in and tap-out with the most popular segment in the system the trip from Victoria to Oxford Circus and vice versa. Waterloo to Canary Wharf is the most frequent during the morning and evening peak with Waterloo and Victoria the two biggest volume hubs in the system.

In understanding cities, origins and destinations of trips, indeed of any flow, is essential for understanding the rationale of the location where those creating the flow are based. One of the problems with smart card data that is orientated to transit systems, such as fixed rail, is that the locations which anchor these infrastructures do not have the same meaning as origins and destinations in terms of work, shopping, residences, schools and so on which generate trips. It is extremely difficult to tie places where people enter such systems to the comprehensive patterns of locations that are described by traditional data. We can quite easily assemble flow matrices and assign trips to network segments such as lines between stations – although the precise paths of travel have to be inferred, but tying these to places of work, residence and so on is difficult. Some headway has been made using smart card data for Singapore (Zhong *et al.* 2014) but the problem is perennial and requires additional data to link points of fixed infrastructure to ultimate origins and destinations.

We have assembled several pictures of transit systems in operation from our Oyster card data. Using shortest path algorithms, Reades (2013) has worked on finding the best routes between stations identified in the data and pieced together actual flows by assigning origin data from tap-ins to the network, then finding the shortest routes on lines linking the origin to the destination. He has produced a computer movie of a typical week from the 2012 data by adding data for several typical weeks – excluding the Olympic Games weeks – thence producing an averaged version which shows the peaks and troughs in the data from Sunday

to Saturday. The weekend days are very different with much less pronounced morning and evening peaks while typical workdays show very distinct morning and evening peaks that in themselves are very different with a small blip in the central area in the late evening (see Figure 3.6).

We are developing several projects using the Oyster card data but so far these tend to examine very different aspects of the city from those that pertain to traditional flow data. The focus is inevitably on questions of disruption and smooth flowing on a fine-scale temporal basis, but we are not able to relate these to links between home and work. We are able of course to examine the variability of the tap-in and tap-out data with respect to the station hubs through two interlocking patterns of entries and exit volumes that reflect two layers of polycentricity which vary through time and are reflected in the peak and off-peak flow patterns. The essential challenge is to tie this to other data, such as activity volumes of employment retailing, residential populations and so on, that come from more traditional sources.

Conclusions and next steps

Big data are never what they seem. The multiple Vs that have become their signature definition do not capture the fact that quite small data when elaborated into their second, third and higher order effects can become big in the sense that conventional techniques and models fail to deal with their extended volumes. Our first illustrations here focus on quite modest data sets and we are conscious that really big data volumes that come from interaction patterns

Clips from the **YouTube** Movie: *Oyster Gives Up Its Pearls*, made by UCL Engineering from Jon Reades's movies of the data

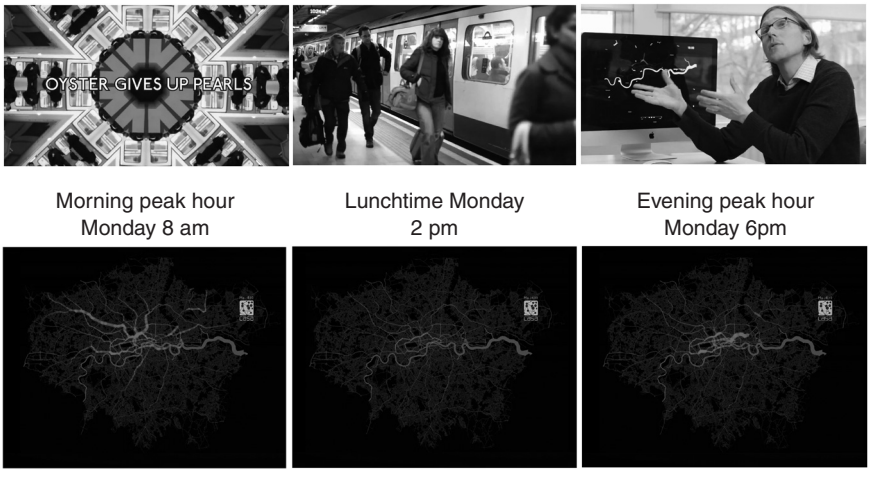

Figure 3.6 Visualizations of the flows on the rail segments during a working day

Movie available at YouTube (www.youtube.com/watch?v=9sAugcb2Qj4)

are hard to measure in terms of their complexity through visualization. The visualization of data in countless ways has proceeded in parallel to the big data revolution, which is focused more on data mining through machine learning and in essence involves iterative techniques for searching for patterns in such data that may or may not have substantive meaning. For example, our illustration of the quality of the fit of our spatial interaction model of journey to work in Greater London (see Figures 3.3 and 3.4), suggests several features of our model and data that are quite counter to one another. In fact, the intensity of points in Figure 3.4 – the fact that a large proportion of points are inside the core of the scatter – probably need to be separated out.

Our continuing work on contemporary big data is taking many forms but so far it is mainly dealing with transit. Data on energy flows and usage in the smart city are not focal as yet, while the analysis of big data associated with social media may well remain in some preliminary form for many years. Representativeness is the key issue, as is meaning in such data, and it is not clear as yet the extent to which these social media data pertain to the social and economic functioning of the city. In another sense, big data are being created or rather extended and conflated through mashups. These kinds of integration are as important as the search for pattern in such data and as the big data revolution proceeds it is increasingly clear that the pronouncements on the end of theory, made so vociferously by commentators such as Anderson (2008), are not being borne out in any sense. The need to approach big data with clear theory has never been more important.

References

Anderson, C. (2008) 'The end of theory: the data deluge makes the scientific method obsolete', *Wired Magazine* 16-07, 23 June, available from: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory [accessed 24 November 2016].

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Batty, M. (2014) 'Can it happen again? Planning support, Lee's requiem and the rise of the smart cities movement', *Environment and Planning B: Planning and Design* 41(3): 388–391.

Batty, M., Hudson-Smith, A., Hugel, S. and Roumpani, F. (2015) 'Visualising data for smart cities', in A. Vesco and F. Ferrero (eds), *Handbook of Research on Social, Economic, and Environmental Sustainability in the Development of Smart Cities*. Hershey, PA: IGI Global, pp. 339–362.

Berry, B.J.L. (1964) 'Approaches to regional analysis: a synthesis', *Annals of the Association of American Geographers* 54: 2–11.

Bos, F. (2011) 'Three centuries of macro-economic statistics', Munich Personal, RePEc Archive (MPRA) Paper No. 35391, available from: http://mpra.ub.uni-muenchen.de/35391/ [accessed 4 December 2016].

Chen, C., Batty, M. and van Vuren, T. (2015) 'Editorial', *Transportation* 42: 537–540.

Dutcher, J. (2014) 'What is big data?', *DataScience Berkeley Blog*, available from: http://datascience.berkeley.edu/what-is-big-data/ [accessed 24 November 2016].

Ferry, G. (2010) *A Computer Called LEO*. New York: Harper Perennial.

Graham-Cumming, J. (2012) 'The great railway caper: big data in 1955', available from: www.youtube.com/watch?v=pcBJfkE5UwU and see http://bigdata.blogweb.casa.ucl.ac.uk/2012/10/03/big-data-problems/ [accessed 24 November 2016].

Han, J., Kamber, M. and Pei, J. (2011) *Data Mining: Concepts and Techniques*. Waltham, MA: Morgan Kauffman.

IBM (n.d.) 'The four V's of big data', available from: www.ibmbigdatahub.com/infographic/four-vs-big-data/ [accessed 24 November 2016].

Kitchin, R., Lauriault, T.P. and McArdle, G. (2014) 'Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards', *Regional Studies, Regional Science* 2(1): 6–28.

Lee, D.B. (1973) 'Requiem for large-scale models', *Journal of the American Institute of Planners* 39: 163–178.

Lowry, I.S. (1964) *A Model of Metropolis*. RM-4035-RC, Santa Monica, CA: The Rand Corporation, available from: www.rand.org/content/dam/rand/pubs/research_memoranda/2006/RM4035.pdf [accessed 24 November 2016].

O'Brien, O., Batty, M., Gray, S., Cheshire, J. and Hudson-Smith, A. (2014) 'On city dashboards and data stores', a paper presented to the Workshop on Big Data and Urban Informatics, 11–12 August, University of Illinois at Chicago, Chicago, IL, available from: http://urbanbigdata.uic.edu/proceedings/ [accessed 24 November 2016].

Reades, J. (2013) 'Pulse of the city', *Vimeo*, available from https://vimeo.com/41760845, original at http://simulacra.blogs.casa.ucl.ac.uk/2011/08/pulse-of-the-city/ [accessed 24 November 2016].

Zhong, C., Arisona, S.M., Huang, X., Batty, M. and Schmitt, G. (2014) 'Detecting the dynamics of urban structure through spatial network analysis', *International Journal of Geographical Information Science* 28(11): 2178–2199.