City Research Online

City, University of London Institutional Repository

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Using treemaps for variable selection in spatio-temporal visualisation

Aidan Slingsby[1]
Jason Dykes[1]
Jo Wood[1]

[1]giCentre, Department of Information Science, City University London, London, U.K.

Correspondence:
Aidan Slingsby, giCentre, Department of Information Science, City University London, Northampton Square, London EC1V 0HB, U.K.
Tel: +44(0)20 7040 0180;
Fax: +44(0)20 7040 8845;
E-mail: a.slingsby@soi.city.ac.uk

Abstract

We demonstrate and reflect upon the use of enhanced treemaps that incorporate spatial and temporal ordering for exploring a large multivariate spatio-temporal data set. The resulting data-dense views summarise and simultaneously present hundreds of space-, time-, and variable-constrained subsets of a large multivariate data set in a structure that facilitates their meaningful comparison and supports visual analysis. Interactive techniques allow localised patterns to be explored and subsets of interest selected and compared with the spatial aggregate. Spatial variation is considered through interactive raster maps and high-resolution local road maps. The techniques are developed in the context of 42.2 million records of vehicular activity in a $98\,km^2$ area of central London and informally evaluated through a design used in the exploratory visualisation of this data set. The main advantages of our technique are the means to simultaneously display hundreds of summaries of the data and to interactively browse hundreds of variable combinations with ordering and symbolism that are consistent and appropriate for space- and time-based variables. These capabilities are difficult to achieve in the case of spatio-temporal data with categorical attributes using existing geovisualisation methods. We acknowledge limitations in the treemap representation but enhance the cognitive plausibility of this popular layout through our two-dimensional ordering algorithm and interactions. Patterns that are expected (e.g. more traffic in central London), interesting (e.g. the spatial and temporal distribution of particular vehicle types) and anomalous (e.g. low speeds on particular road sections) are detected at various scales and locations using the approach. In many cases, anomalies identify biases that may have implications for future use of the data set for analyses and applications. Ordered treemaps appear to have potential as interactive interfaces for variable selection in spatio-temporal visualisation.
Information Visualization (2008) **7,** 210–224. doi:10.1057/palgrave.ivs.9500185

Introduction

Large multivariate spatio-temporal data sets – for example, traffic flow data or mobile telephone logs – are likely to contain structure and patterns that provide useful information about characteristics of the measured phenomena. The identification and comparison of such patterns may assist in understanding these phenomena and may be used for a number of purposes including research, gaining competitive advantage, planning and other operational tasks. The complexity that arises from the interactions among the spatial, temporal and attribute aspects in such data sets[1,2] and the imprecise goals that are often associated with their initial exploration[3] makes the identification of patterns and structure challenging.[1–3] Information visualisation and geovisualisation techniques are an increasingly
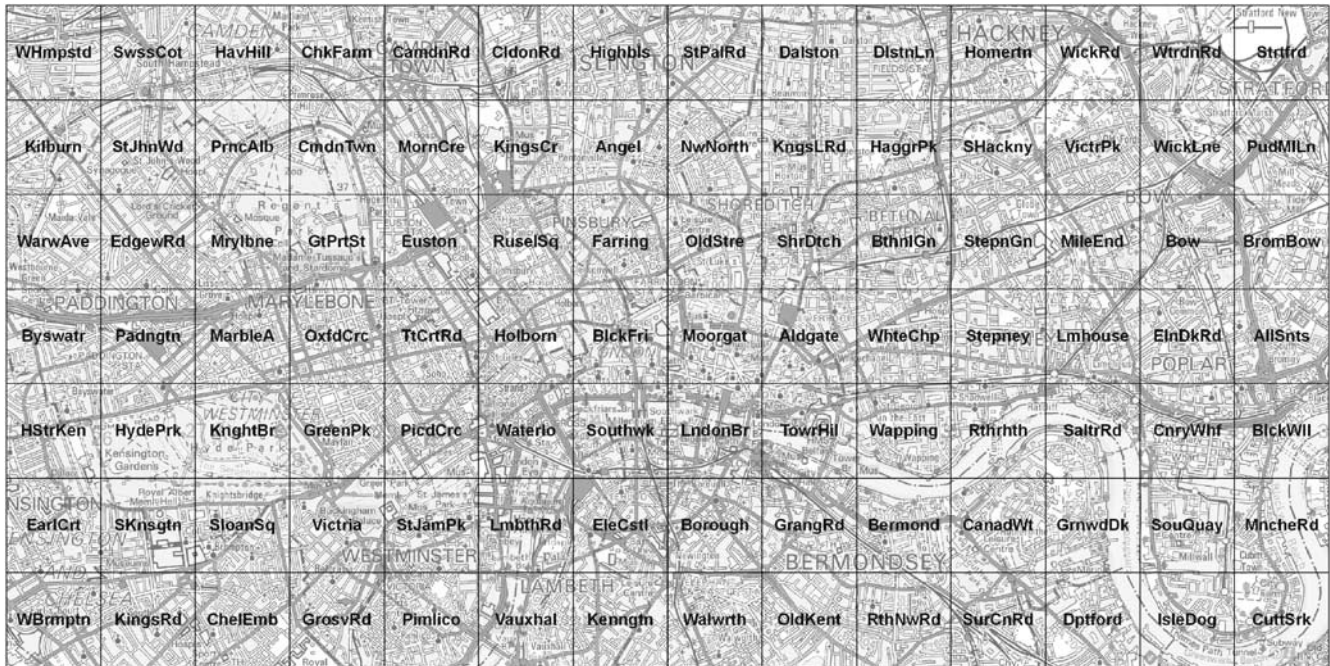
Figure 1 The central London area from which the GPS points were collected, showing labelled 1 km grid squares used as part of this study. *Base map*: © *Crown Copyright/database right 2008. An Ordnance Survey/EDINA supplied service.*

important means of assisting in the process of finding patterns and structure in large multivariate spatio-temporal data sets.[4–7] One approach to this process involves comparing *variable-constrained subsets* of a data set and selecting interesting variable combinations for further inspection. For multivariate spatio-temporal data sets, this can amount to hundreds or thousands of possible subsets. Data-mining techniques can help reduce this search space by selecting subsets worthy of further comparison and their combination with visualisation techniques is characteristic of visual analytics.[8] We have developed a novel visual approach that uses treemaps with spatial and temporal ordering to simultaneously present thousands of summaries of variable-constrained subsets of a 42.2 million record data set. These serve as rich *data-dense overviews* whose systematically ordered nature may facilitate the broad comparison of subsets. Interactive techniques are then used to support the selection of particular subsets of interest. We present a design in which these views can be further explored using alternative treemap views, raster maps and road maps. We demonstrate these methods by visually exploring the data set, identifying patterns and structure, and informally evaluating the techniques used through informed reflection.

Data

The London-based courier company eCourier[9] collected 42.2 million GPS points from delivery vehicles (an average of 48 vehicles per day) at approximately 10-s intervals

between June 2006 and May 2007 in a $98 \, km^2$ area of Central London (Figure 1). Each record contains the vehicle's position, speed, vehicle type (van, large van, motorbike, large motorbike or bicycle) and the time at which it was collected.

This data set is interesting for a number of reasons. Firstly, its analysis may be of specific value to the courier company concerned[10] to help optimise vehicle allocation, scheduling and routing. Secondly, the techniques may also be of more general interest; for example transport authorities assess patterns of traffic flow to help set policy to reduce congestion.[11,12] Thirdly, it typifies a trend whereby large data sets such as those that are volunteered or derived from computer logs are released through open APIs.[13] This is increasing opportunities for (geo)visual analysis[14] and is fuelling the emerging field of visual analytics.[8,15] In this context, visual data exploration techniques can be used to identify patterns of interest that may relate to significant characteristics of the phenomenon under study. Importantly, they may also help draw attention to biases and data quality issues in large informal data sets of unknown quality.

Visualisation challenges

Large spatio-temporal multivariate data sets pose substantial visualisation challenges in terms of data complexity – both the interactions among spatial, temporal and attribute aspects[2] and the complex relationships between the data collected and the phenomena under

consideration. Large data sets (millions of records) are prone to the trade-off between the potential for visual occlusion caused by overplotting and the loss of resolution inherent in highly aggregated summaries.[1] Memory and computation overheads are significant when processing large data sets, which can make interactive querying to produce the 'instant' response times required to support visualisation[2] difficult. This applies to the eCourier data set through the following characteristics:

- *Large size*: The 42.2 million records pose challenges for providing interactivity.
- *High spatial density*: The 1 km grid squares contain between 0.3 and 227 points per metre of road (mean = 20) positioned on 1831 km of road network comprising 28,838 road segments. Visual occlusion is a problem for displaying such data making the use of colour symbolism on the road geometry ineffective at the global scale.
- *Multivariate nature*: A number of variables can be derived from the data. We use hour of the day, day of the week and 1 km^2 grid cells as categorical spatial and temporal variables in our analysis, alongside the recorded vehicle type. Ignoring space for the moment, we can select from 1199 possible combinations of single values of the three categorised variables (24 hours of the day, 7 days of the week, 5 vehicle types; $5 + 7 + 24 + (5 \times 7) + (5 \times 24) + (7 \times 24) + (5 \times 7 \times 24)$). This number rises markedly when we consider geographic subsets – the 98 possible grid squares for example. Existing techniques such as small multiples or animation are not able to give access to all these subsets simultaneously.
- *Spatio-temporal aspects*: Spatial and temporal data have inherent ordering essential for comparison, interpretation and assimilation. These aspects must be graphically represented so that subsets can be compared in their spatial and temporal contexts.

Approach

The challenge is to develop views and interactions that provide access to information about the relationships between these subsets and their spatio-temporal characteristics in a manner that aids comparison and assimilation. Our approach has two elements – broadly following Schneiderman's[16] 'information-seeking mantra'. Some innovation was required in order to address the challenges outlined above in our efforts to explore the eCourier data set.

- *Overview* – Treemaps with spatial and temporal ordering simultaneously provide rich data-dense summaries of hundreds or thousands of subsets of the data set, using consistent ordering that reflect its spatio-temporal nature, without visual occlusion.
- *Zoom, filter and details on demand* – An interactive design for exploration through which variable-constrained

subsets of interest can be selected for inspection as:

  (a) raster maps for the entire area – no visual occlusion;
  (b) road maps for individual grid squares – an appropriate scale for displaying road segments such that the symbology is discernable;
  (c) treemaps for comparing local data with global summaries.

We use four visual techniques in our design and develop links between them. The first is used for an overview and the latter three for zooming, filtering and obtaining details on demand. Figure 5 shows a screenshot of our design – an interactive prototype – that contains:

- *Spatial treemaps*, with *fixed-size* nodes and spatial and temporal ordering – coloured by traffic volume and by speed. These show overall spatio-temporal patterns in vehicle use. We term these layouts 'maptrees' where the top-level of the hierarchy is a spatial unit (Figure 4).
- *Interactive attribute treemaps* (top left of Figure 5) sized by *global* traffic volume, coloured by *local* (1 km^2 grid squares) traffic volume or speed allowing the comparison of global characteristics with local characteristics of a subset. Interesting subsets of data represented by nodes and leaves in the tree can be *selected* (e.g. vans on Monday) for display as raster maps (to show the spatial distribution of the subset's traffic volume or average speed) and road maps (for individual grid squares).
- *Interactive raster maps* (bottom of Figure 5) show the spatial variation of subsets selected in the interactive attribute treemap. They facilitate the selection of individual grid squares, allowing the subset summary to be compared with the global traffic volume (in the interactive treemaps) and mapped as a road map.
- *Road maps* (top right of Figure 5) are shown for the 1 km^2 grid cells selected in the raster map and map the traffic volume or average speed for each road segment. Brushing the road segments numerically displays details on demand – the number of data points and average speed for that segment.

This approach attempts to address the challenges posed by large multivariate spatio-temporal data sets using visual encodings that do not exhibit visual occlusion (treemaps and raster maps), give access to coarse and fine-grained aggregates (subsets) and support their visual comparison. It is scalable due to the use of pregenerated summaries at hundreds or thousands of levels of aggregation. Our treemap algorithm is implemented as a Java application[20] with output in SVG and a number of image formats. The designs are encoded in SVG and interaction is provided using the DOM with JavaScript.

Treemaps for showing multivariate data

Treemaps display hierarchical data[17] by recursively and exhaustively subdividing space at each level in a
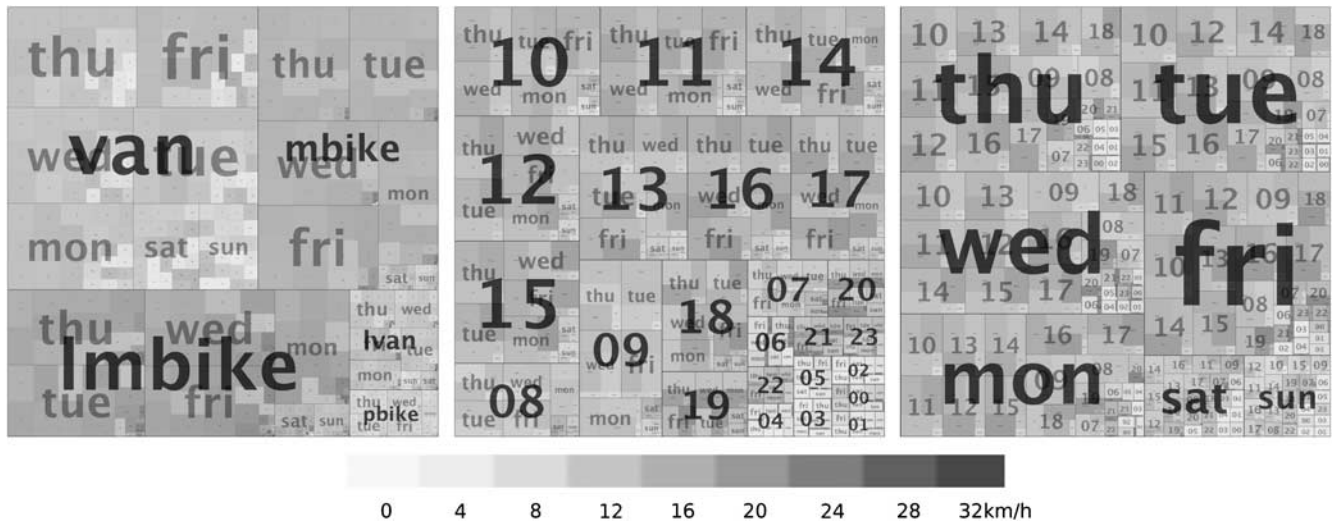
Figure 2 Squarified treemaps summarising relative traffic volumes (size) and average speed (colour; km/h) for multiple subsets based on the three false hierarchies. From left to right these are: type–day–hour, hour–day–type and day–hour–type. Average speed is symbolised using ColorBrewer 'Oranges'[22] throughout this paper (all images are in colour in the online version).

hierarchy, using a form of dimensional stacking.[18] The data-rich displays that result make efficient use of space and so are amenable to showing large hierarchical data sets. Each item in the hierarchy can be considered a node. We distinguish between leaf nodes that have no further subdivision and branch nodes that do. Information can be represented by the *size* of nodes, the *colour* of leaf nodes, the *spatial order* (or more generally, layout) of nodes and a *label* for each node.

Hierarchical visual representations can be used to display multivariate data.[18–20] They require variables to be categorical or transformed into such (e.g. date transformed into day of the week) and structured into a false hierarchy (e.g. vehicle type, day of the week, hour of the day). By using combinations of categorical variable values to define subsets (e.g. vans on Fridays at 06:00–07:00; vans on Fridays; vans), potentially hundreds of hierarchical variable-constrained subsets can be defined on this basis. In Figure 2 (left) such subsets are visually represented in an ordered squarified treemap.[21,20] In this case, node size and order (from top left to bottom right) is used to show relative traffic volume and colour is used to show average speed. The treemap reveals that vans comprise the largest share of traffic in the eCourier data set and bicycles ('pbike') the smallest share. Weekdays are the busiest, with weekend traffic approximately half that of weekdays. Large motorbikes ('lmbike') have the highest average speeds and these are associated with weekday records. The different spatial arrangements allow us to see relationships between variables in the context of the numbers of records in each combination. For example, vans and large vans have slower more variable speeds (left), less traffic occurs in the middle of the night and this is slow (centre), nighttime traffic is similar in volume throughout the week and Sunday is the day with least

eCourier traffic yet average speeds tend to be slow (right). The ability to detect and compare such suggested patterns is a key objective of overview graphics.

One of the limitations of using hierarchies in this way is that not all possible subsets are discernable as continuous areas in any one treemap. Where there is no natural variable hierarchy, multiple equally valid hierarchies are possible.[23] Figure 2 shows the same data structured in three alternative hierarchies, illustrating some limitations of using single treemaps to present variable-constrained subsets:

- Not all possible subsets are available through a single hierarchy.
- Colour can only be applied to represent the values of *leaf nodes*.
- It is difficult to compare nodes that are in different branches of a hierarchy. For example, we cannot easily distinguish the number of 'vans at 10' in Figure 2 (left) because the 'van' leaf node and the '10' leaf node are not spatially contiguous as they would be in the type–hour–day and the hour–type–day hierarchies.

Some of these shortcomings can be addressed to an extent by switching the hierarchy. Figure 2 reveals that there is more daytime traffic (middle), that this is consistent across the week (middle) but less consistent between vehicle types (right). The high consistency of colour in the treemap on the right shows that traffic speed has a higher dependence on the vehicle type than on the hour of the day.

The maximum possible depth of the hierarchy depends on the number and sizes of leaf nodes and the size and resolution of the screen. In practice, hierarchical levels beyond level 4 become difficult to resolve and we would

Figure 3   Squarified treemaps showing the same data as in Figure 2 but with fixed leaf sizes and temporal ordering (midnight is at the top left), coloured by traffic volume (purple; logarithmic scale; numbers of points) and average speed (orange; grey where there are no vehicles; km/h). Traffic volume is coloured using ColorBrewer 'Purples'[22] throughout this paper.

recommend using multiple treemaps where this is the case.

Colour relates to values at *leaf nodes*, but consistency of colour can be used to infer range and variability *within branches*. Figure 2 (left) shows that large motorbikes generally have a higher average speed than vans and that there is more variability in speed of van traffic than motorbike traffic. For this reason, it is useful to be able to change the depth of the hierarchy and switch hierarchy through an interactive design that involves multiple alternative treemaps.

Treemap layout and order

Information visualisation techniques project data onto Euclidean two-dimensional planes for display, a process known as spatialisation.[24] Although geovisualisation techniques can use well-established cartographic coordinate systems, the way in which non-spatial data should be spatialised is less clear-cut. Skupin and Fabrikant[25] reasonably argue for the use of consistent spatial metaphors in layouts, such as Tobler's so-called 'First Law of Geography'[26] where proximity can be associated with relatedness, in order to improve cognitive plausibility.

The recursive subdivision of space at each level in the treemap hierarchy has the effect of isolating layouts within each level, leading to discontinuities between hierarchies and abrupt contraventions of Tobler's 'First Law' . For example, the relative position of 'Friday vans' and 'Friday motorbikes' is arbitrary in Figure 2 because they are ordered independently of each other, making comparison of Friday for different vehicle types difficult. Concerns about the cognitive load imposed on the user that such discontinuities and inconsistencies may have are expressed in the literature.[25] We acknowl-

edge and address some of these concerns with our enhanced treemaps that use consistent layout, appropriate ordering and interaction, and also argue that some concerns relating to treemap usability[27–29] involve very different tasks and contexts to those under consideration here.

Node size is an appropriate ordering criterion for comparing magnitudes (Figure 2). However, where data sets contain spatial and temporal structure, categories and category combinations may have inherent orderings in time and space. Where categories are ordered in one dimension (e.g. 'day of week') we apply an 'ordered squarified' algorithm,[20] ordering from top left to bottom right. Where categories are ordered in two dimensions (e.g. spatial subsets) we use spatial ordering.[20] Both these techniques produce consistent ordering within and between hierarchies and thus more cognitively plausible layouts that support the comparison of overall patterns across and within hierarchies. These techniques mean that while hierarchical information is maintained in the layout, spatial relationships in the treemap relate more closely to one or two-dimensional relationships in our variables. We argue that these enhancements reduce the cognitive load and increase the plausibility of this particular spatialisation.

Temporal ordering

For temporally consistent ordering, we use treemap leaves of constant size. While we lose one information carrying dimension, by using size consistently to give every subset equal prominence, we gain another – order. Additionally, low volumes of traffic – which may be as worthy of further exploration as high volumes – are more easily detected. Figure 3 shows the same treemaps presented in

Figure 2 but with fixed leaf sizes and temporal ordering (midnight is at the top left; vehicle type ordering is arbitrary but consistent). Since we have lost the property of size for conveying numerical values we need a second treemap – coloured by traffic volume (purple). The traffic volume treemap shows striking temporal patterns. The repeated diagonals are expected, showing that most traffic occurs during daylight hours on weekdays; however, some patterns are perhaps less expected. Van traffic appears to be heavy at all times, with a large increase in daytime traffic in comparison to night traffic (see logarithmic scale bar). The speed treemap shows that patterns of average speed are not so strongly correlated with time, but that van traffic, motorbike traffic and large motorbike Saturday traffic tends to be slower at night, a surprising finding that may be worthy of further investigation – are there spatial patterns to this trend for example?

While adjacencies of leaf nodes at the boundaries *between* branches are fairly arbitrary, temporal ordering *within* branches introduces an ordering consistency that enables temporal patterns to be identified even if the detail of the leaves is not visible.

Spatial ordering

Treemaps usually employ one-dimensional ordering in two-dimensional space. This is the case in Figures 2 and 3 for the levels of the hierarchy relating to day and hour. Where spatial data are involved, two-dimensional ordering can be used, resulting in spatial treemaps.[20] By subsetting the data set into geographic units (such as the $1 km^2$ square shown in Figure 1), inserting the spatial subsets into the base of the variable hierarchy (grid square name and location), fixing the node size and using the correct aspect ratio, a spatially ordered treemap can be produced. This might be termed a 'maptree', because as shown in Figure 4, this is effectively a (geographical) map of localised versions of the treemap shown in Figure 3. The consistent ordering at the appropriate levels of the hierarchy can be used to draw attention to spatial and temporal patterns across the entire data set. For example, it is not surprising to note that the highest traffic volumes are around the centre, but there are also high volumes of traffic at certain times of day in the east. Temporal patterns for each grid square can be seen; for example, grid squares in the centre and towards the southwest have higher volumes of van traffic at all times (upper left of each grid square) and high daytime volumes of large motorbike use (lower right of each spatial square), but nighttime and weekend van traffic is much lower in the east. Bicycle traffic (top right of each spatial square) is only found in the centre and towards the northwest and motorbike traffic is almost non-existent in the northeast. The speed treemap shows lower speeds (as expected) in the centre, but isolated grid squares can be picked out containing consistently high average speeds. The fast squares in northeast and west London have high speeds associated with vans (top left of each spatial square) and

large motorbikes (bottom right of each spatial square). These high speeds are associated with main roads (M40 in west; A12 in northeast; see Figure 1). In the south and just north of the centre, it is only large motorbike traffic with particularly high average speeds. The initial visualisation suggests that these combinations of space, time and attribute may warrant further investigation.

Although we focus on false attribute hierarchies here, inherent hierarchies relating to different granularities of space and time are a common consideration (e.g. counties within countries). The effect of spatial granularity on statistical aggregates can be explored using spatial treemaps as hierarchical cartograms,[20] by creating hierarchies from coarse to fine granularities; for example, where colour would show values of leaf nodes (finest granularity) and size represents relative values within and across whole hierarchies (if the value is additive through spatial granularities – as is traffic volume). However, because we have chosen to fix node size, little benefit would be gained from looking at different spatial granularities in our 'maptrees', other than through interactively changing the hierarchy depth. Instead, we choose the two spatial granularities, $1 km^2$ grid squares (we found this spatial resolution to be helpful for the maptrees and raster maps and they correspond to National Grid mapping squares) and road segments.

Interactive methods

The overview techniques have resulted in the identification of patterns and related ideas that warrant further exploration. To support this activity through an iterative processes of overview, zoom, filter and details-on-demand[16] we use a series of interactive techniques (listed in the next subsection) to link treemaps, spatial treemaps, raster maps and road maps. Doing so in response to ideas generated through the overview treemaps and maptrees allows us to examine interesting subsets of data in detail and at a higher resolution.

A screenshot of the iteratively developing design used in our analysis is shown in Figure 5. It employs open technologies for high-level scripting, including HTML, SVG, JavaScript and CSS in a manner that has evolved as our data exploration has progressed. It contains a series of novel aspects and reveals structure in the eCourier data set for various variable combination subsets at a number of scales.

Hierarchy switching and changing hierarchy depth

We have drawn attention to the need to switch hierarchy and change the hierarchy depth. Our design allows different levels of the attribute hierarchy to be selected. In Figure 5, the interactive treemap is shown at level 1 – aggregating by transport type and in this example enabling all van traffic to be selected. In Figure 6 (top), all three levels of the hierarchy are shown, each with a local colour scheme showing the noise associated with
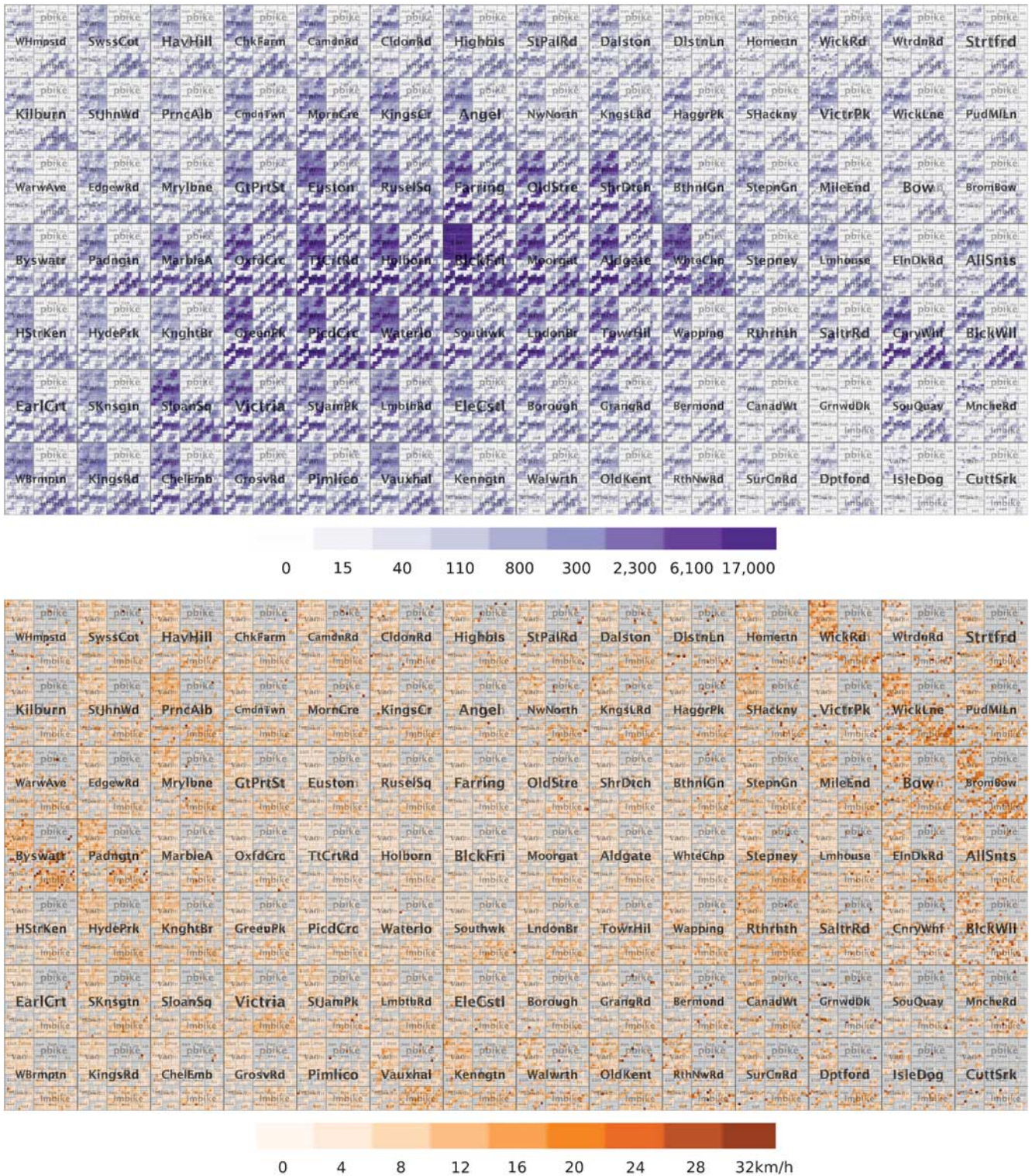
Figure 4 Spatial treemaps or 'maptrees' coloured by traffic volume (purple; logarithmic scale) and average speed (orange; linear scale), using the false hierarchy: grid–type–day–hour. The size of leaves is fixed and temporally ordered within grid squares, with grid squares in their true geographical configuration.
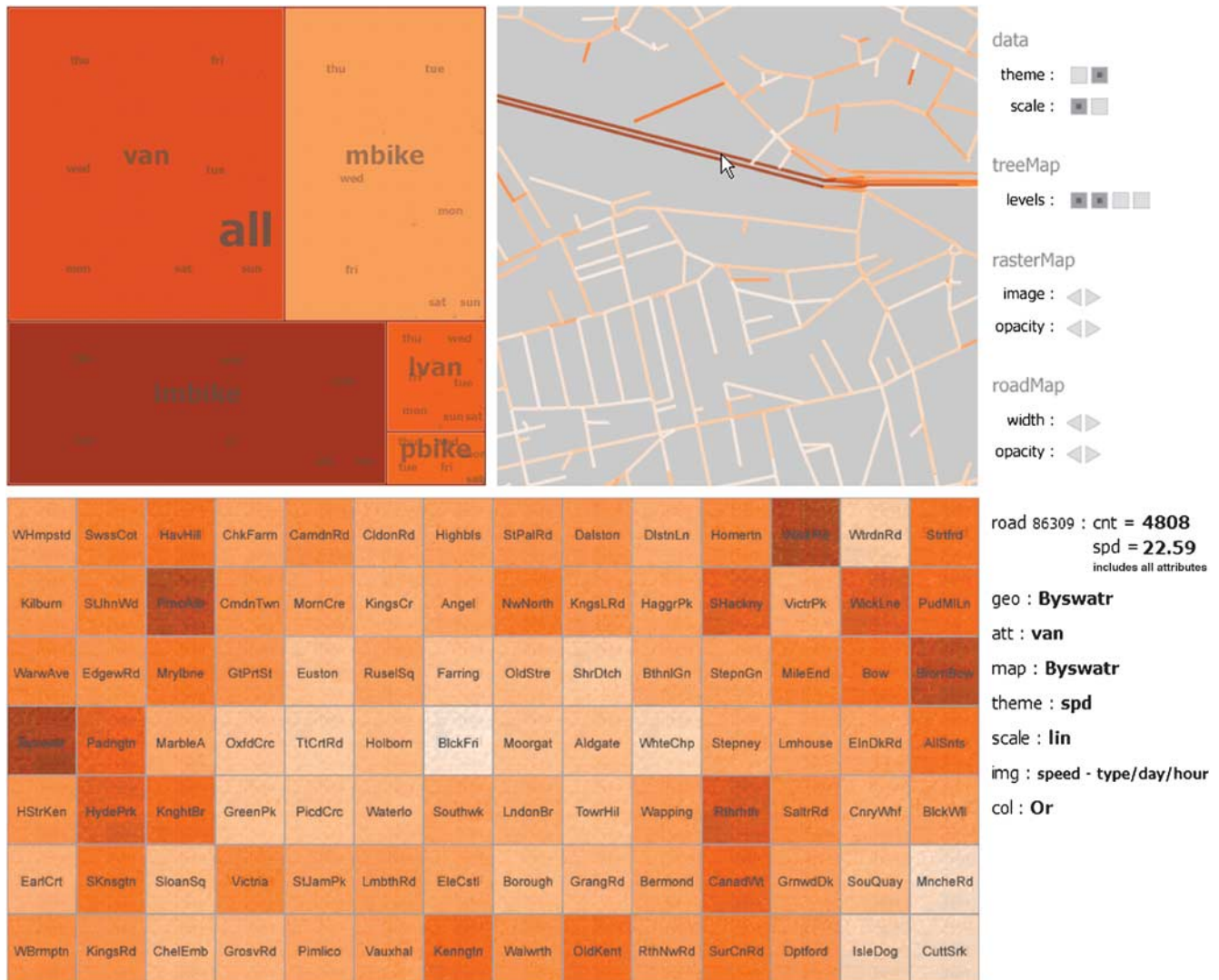
Figure 5   Design contains an *interactive treemap* (top left; sized by global traffic volume, coloured by local traffic speed for transport type in Bayswater), *road map* (top right; for Bayswater, coloured by van traffic; the fast road is the M40 main road, other roads are residential side streets), *raster map* (bottom; coloured by van traffic) and some *controls and detail derived through brushing with the cursor* (right). The van traffic variable was selected from the interactive treemap by changing its depth to one variable ('levels' radiobutton) allowing it to be selected. Bayswater was selected from the *raster map*. Moving the mouse over a road displays the number of GPS points ('cnt') and average speed ('spd') on the right. *Road maps are derived from the ITN layer of OS MasterMap,* © *Crown Copyright/database right 2008. An Ordnance Survey/EDINA supplied service.*

subsets containing small amounts of data. The technologies employed here also enable us to load alternative treemaps as and when required.

Treemaps for comparing global and local patterns

Treemaps offer the properties of colour, size, labelling and order to convey data values and other information. Although the dependency between size and order is problematic when using treemaps for global summaries (positional inconsistencies between hierarchies), size and

colour can be usefully used for comparing global with local patterns. In Figure 6 (bottom), we use size and colour to represent *global* and *local* traffic volume, respectively, for a variable subset and a grid square. Local colouring is selected by clicking cells in the raster map. Where large nodes have dark shading or small nodes have light shading, the *global* (all traffic for the whole area) and *local* (traffic volume for the selected subset in the selected grid square) patterns are most similar. The treemaps in Figure 6 (top) are coloured by average speed, so that large dark nodes represent the situation where both *global* traffic
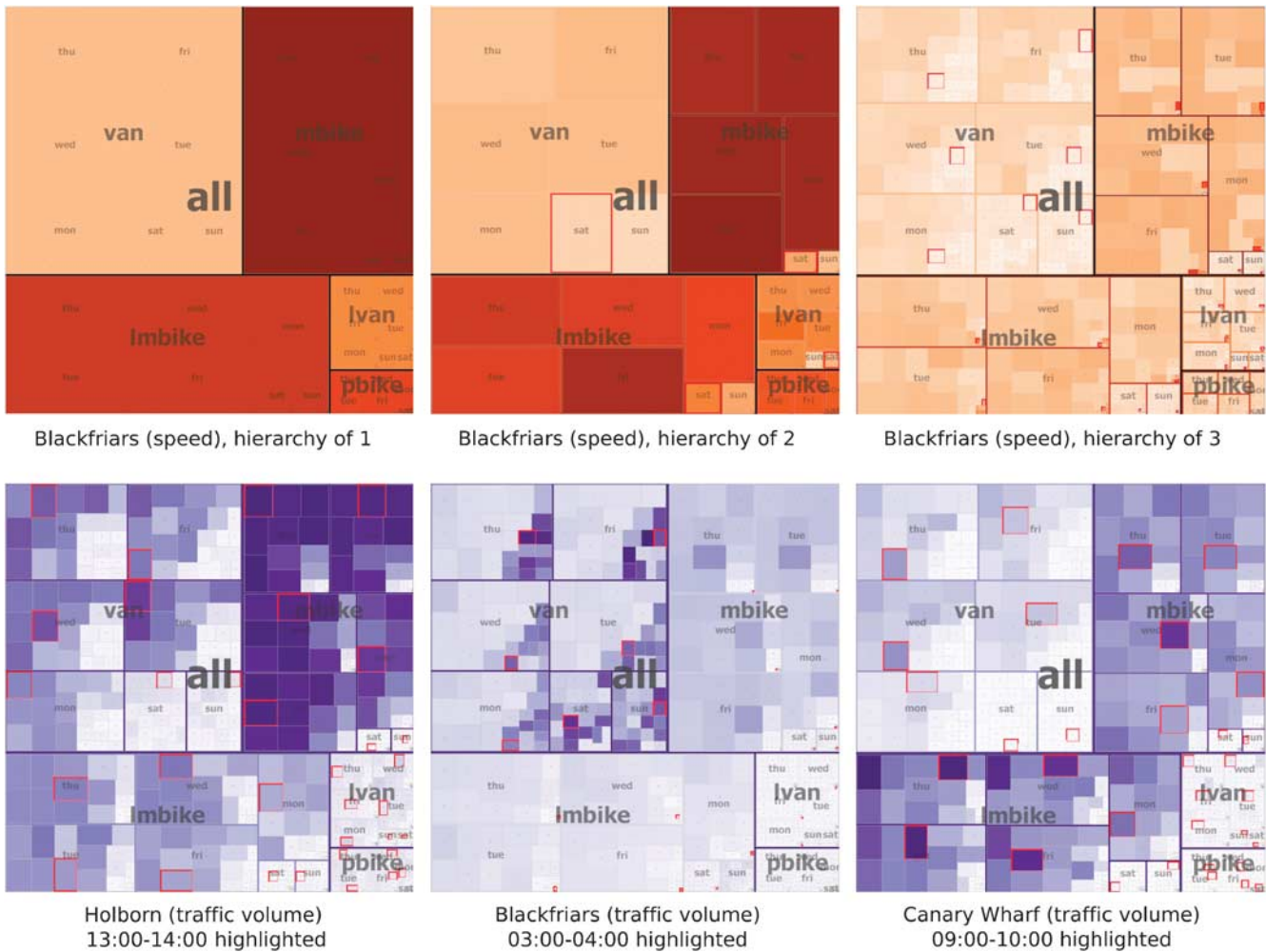
Blackfriars (speed), hierarchy of 1

Blackfriars (speed), hierarchy of 2

Blackfriars (speed), hierarchy of 3

Holborn (traffic volume)
13:00–14:00 highlighted

Blackfriars (traffic volume)
03:00–04:00 highlighted

Canary Wharf (traffic volume)
09:00–10:00 highlighted

Figure 6   Top: Interactive treemaps of the type–day–hour hierarchy at three different depths, whose nodes are *sized by global traffic volume* and coloured by *local average speed* with local colour schemes. The lighter colours seen on the right are due to outlier high speeds for motorbikes (highlighted). The highlighting shows that these high speeds are at consistent times of day through the week. Bottom: Interactive treemaps of the full type–day–hour hierarchy, whose nodes are *sized by global traffic volume* and coloured with linear *colour scheme according to local traffic volume* as selected through the raster map for Holborn, Blackfriars and Canary Wharf.

volumes and *local* average speeds are high. The problem of positional inconsistency between nodes in different branches of the hierarchy can be addressed through hierarchy switching or by using interactive brushing whereby equivalent nodes are highlighted across a particular hierarchy as shown in Figure 6. This latter technique provides clear visual clues to address an issue that has been cited as problematic in terms of cognitive plausibility.[25]

These interactions allow us to use the raster map to filter by geography and can draw attention to local variations between the variable combinations displayed in the treemap. For example, the logarithmic colour scaling used for traffic volume in Figure 4 (top) appears to show that Blackfriars has a very large volume of van traffic

at all times. However, the local colour scaling in Figure 6 (bottom middle) shows that traffic volume is heavily skewed towards times at which global traffic volumes are usually low (small dark nodes–03:00–04:00 is highlighted). In contrast, Figure 6 (bottom left) suggests that the temporal distribution of local van traffic in Holborn (Blackfriars' western adjacent neighbour) is similar to the global pattern (large dark nodes). It also indicates that there is relatively more motorbike (top right of treemap) than van traffic (top left) in Holborn than globally. Figure 6 (bottom right) reveals another interesting pattern; that there are particularly high volumes of traffic at 09:00–10:00 (highlighted) on weekdays in Canary Wharf and that motorbike delivery is key in this area. Interacting with the maps and treemaps can help us discover

the extent to which this pattern follows 'Tobler's Law' or can be considered a local outlier.

Raster and road maps

Raster and road maps are used to consider spatial patterns of particular variable combinations. The high number (28,838) of road segments in central London, their variation in length and dense geographical arrangement make it difficult to discern all roads in a single overview, let alone represent additional attribute information for visualisation through colour. Broad spatial patterns in filtered subsets of large data sets can be considered and compared using traditional raster maps. Where we need to inspect the geography of a grid square in detail, we use generalised road maps that summarise traffic volume and speed on particular segments. A 38% random sample of GPS points was snapped to nearest road segments in the examples provided here.

Maps for any variable combination can be interactively selected by clicking nodes in the treemaps in our design in Figure 5; so for example, clicking the 'van' in the treemap will shade the raster map according to 'van' traffic. Changing the depth of the treemap hierarchy would enable us to generate raster maps based on the values of more than one variable, for example motorbikes on Thursday at 13:00–14:00 is shown in Figure 7D (left). Selecting 'GrnwdDk' will produce a road map (Figure 7D, right) and colour the treemap according to the local situation there. These views allow us to study the spatial structure of this subset and compare with the global situation. Figure 7 (A, B and C; left) show raster maps of all van traffic (the top-level node in our treemap). Blackfriars dominates in terms of van numbers and has an unusually low average speed. Our interactive techniques allow us to subsequently select the Blackfriars grid square and view its local treemap and map its traffic summarised by road segment (Figure 7A, B and C; right). The logarithmic scale hides the large variation at the upper end of the scale, but when a local linear scale is used (Figure 7B, right), it becomes clear that the majority of vehicles are found on one no-through road. Since this clearly does not represent through traffic, this is consistent with the anomalously low-speed observed.

Inspection of individual values and comparison with global trends

Our design provides important details on demand. Values associated with individual symbols are displayed when the symbols are touched enabling us to compare average speed with traffic volume – effectively the sample size. This is important where outliers may be caused by particularly small samples. Figure 7 (bottom) shows the average speeds for motorbikes on Sunday at 13:00–14:00, representing a small subset of the data. The raster map shows that traffic in the 'GrnwdDk' (6,12) grid square has a particularly high average speed. Inspecting and

querying the road map shows that this high speed is associated with one road segment with a single GPS point. This outlier is likely to have a significant effect on the graphics.

Maptrees can be compared with the raster map so that the data-dense overviews can inform the visual data exploration process (see semi-opaque maptree and 'opacity' control in Figure 5). Various images of these overviews can be loaded (see 'image' control in Figure 5) including maptrees of standard deviation and coefficient of variation of speed, enabling us to account for outliers such as the 'GrnwdDk' road segment.

These linked views and coordinated interactive techniques allow us to identify specific instances and broad trends through aggregated summaries at various levels as well as detailed information about data points at different resolutions.

Discussion

Visualisation challenges

Large, multivariate spatio-temporal data sets pose problems for the design and implementation of effective visual data exploration systems. Keim *et al*.[1] noted that many visual techniques either aggregate and summarise the data to a high degree (e.g. barcharts) or suffer from the visual occlusion of overlapping data points. We try to address these potential weaknesses with our combination of treemaps (that simultaneously show multiple aggregates ranging in aggregation from high to low, exhaustively tessellating space), raster maps (spatial aggregates that exhaustively tessellate space), maptrees (that combine these two approaches) and localised generalised road maps (aggregated to a finer spatial resolution at an appropriate spatial for the density of the road network).

If the only means of visual encoding were the traditional cartographic techniques of raster maps for large areas and roadmaps for localised areas, it would be difficult to compare summaries of so many data subsets. Techniques such as small multiples, animation or interactive selection could be used, but these would significantly restrict the number of comparisons that could be made. The simultaneous display of summaries of all these subsets through false hierarchies with spatial and temporal ordering is a novel use of the treemap that enables broad comparisons between subsets to be made and has enabled us to identify characteristics that are both expected and surprising. Adding interactivity to link treemap overviews to the more traditional techniques of raster maps and roadmap, allows interesting subsets to be inspected in detail. This provides particularly rich overviews and the opportunity to zoom, filter and obtain details on demand.

Treemap issues

Treemaps are data-dense and efficient data visualisation techniques for hierarchical data. Their popularity is due in part to the intuitive recursive division used to represent

Volume of all van traffic (logarithmic scale) for whole area and Blackfriars

Volume of all van traffic (linear scale) for whole area and Blackfriars

Average speed for all van traffic (linear scale) for whole area and Blackfriars

Average speed for motorbike traffic on Thursday at 1300 (linear scale) for whole area and "GrnwdDk"

Figure 7 Raster maps and road maps. Road maps are derived from the ITN layer of OS MasterMap, © Crown Copyright/database right 2008. An Ordnance Survey/EDINA supplied service.
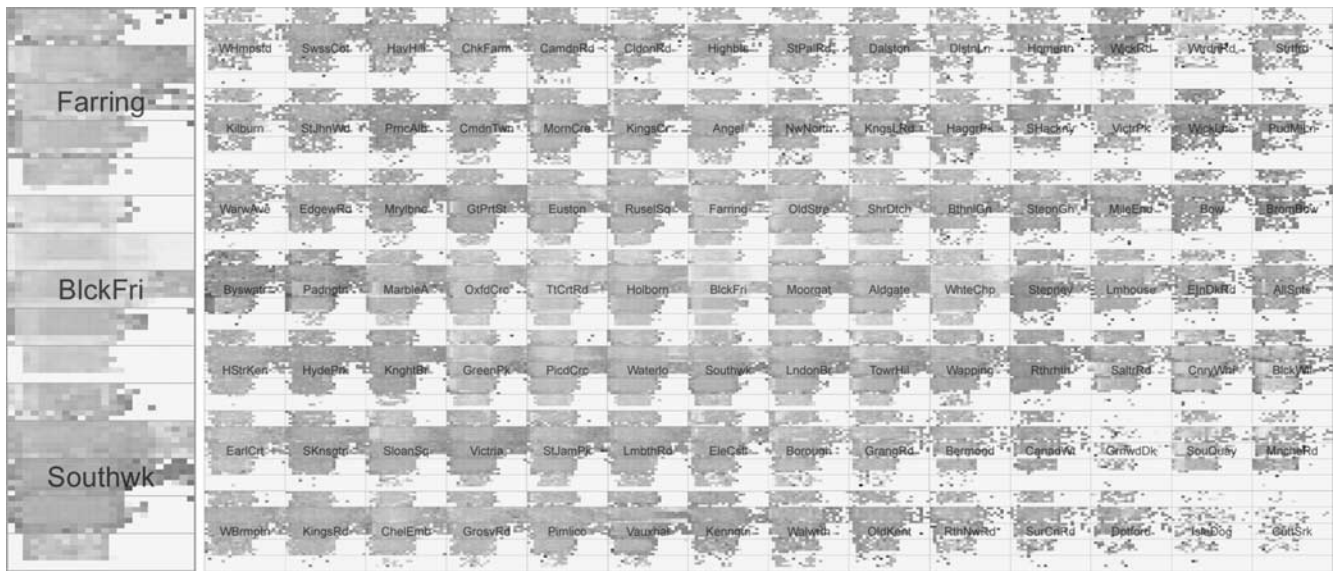
Figure 8 Alternative 'calendar views' coloured by average speed in which vehicle types are arranged in five thick horizontal bars, within which are seven rows (Monday at top) and 24 columns (06:00 on left). Enlarged versions of three grid squares are shown on the left.

hierarchical data in a map-like fashion and the relative ease of implementation.[25] However, a number of known limitations are well reported in the literature. Skupin and Fabrikant[25] suggest that the cognitive plausibility of visualisation techniques is likely to be improved by using cartographic metaphors, but do not define criteria for assessing this. The various extensions and modifications we have documented address some of these concerns and include the incorporation of cartography's primary metaphor – two-dimensional ordering – within the treemap layout.

The eCourier data set has no natural hierarchy and transforming data into an arbitrary false hierarchy for visual analysis and variable selection as proposed here may introduce unhelpful or misleading artefacts. We address these problems by switching between hierarchies (Figure 3), and designing interactive techniques for changing the depth of the hierarchy and providing visually relating discontinuous nodes that represent related categories across our hierarchies (Figure 6).

The 'ordered squarified' and 'spatially ordered' algorithms[20] have the potential to produce more cognitively plausible treemaps. In the section Temporal ordering, we demonstrated that by fixing the node size, we can ensure a consistency in ordering that allows spatial and temporal patterns in categorical data to be identified through repeated patterns (Figure 4). However, while the ordered hierarchical representation of day and hour allowed us to identify temporal patterns, it might be considered to be stretching cognitive plausibility somewhat as the variables are not unrelated. They might be visually associated more appropriately and effectively through a non-hierarchical representation. Figure 8 shows

an alternative approach to spatialising time in which the dimensions of the plane are used to represent the related temporal orders – day of week and hour of day. Our spatial ordering method[20] is then used to locate the leaves that are positioned in terms of hour of day (x) and day of week (y). The higher level of the hierarchy ('type') is consistently ordered in a linear fashion using a standard 'slice and dice' approach that provides elongated spaces for our hour × day ordering. The result is a calendar-like alternative that provides a synoptic view of change over time and allows variation to be considered within each vehicle type (five thick horizontal rows) by day of the week (seven thin horizontal rows) and hour of the day (24 vertical columns). Daily and weekly patterns can be discerned through this ordering, which supports the comparison of hourly and daily cycles and variations across vehicle types – for example, the slow weekends and night times are apparent. Note also the high-speed motorbike speeds in Blackfriars at 06:00, as discussed in relation to Figure 6 (fourth row, first column of enlarged 'Blackfriars' cell to left of Figure 8). While these views use space to show discrete ordered phenomena in each dimension they still require cognitive effort to relate a cyclical phenomenon (the passage of time) to a linear representation.

Figure 7 demonstrates the need to vary colour schemes by showing the impact of varying between logarithmic and linear scales. With logarithmic colour scaling, patterns across the entire area can be exposed, but the impact of outliers at the upper end of the scale is largely ignored. Sufficient flexibility is required to allow rescaling of colours so that both general trends and extreme values can be represented. For example, using a local road map with a linear scale, we were able to identify that a very

large proportion of vans producing GPS records were essentially static on a single no-through road. Our implementation allows us to vary and scale the colour schemes used in the maps and treemaps.

Treemaps with a hierarchy depth of 3 or more had leaves that were difficult to resolve and label. Treemaps with fixed-size nodes had a consistency of node order (e.g., vans always at the top left; daytime hours as diagonal stripes), which once established can be picked out even where node labels are too small to be legible. Discontinuities between nodes in different hierarchies made it difficult to compare individual values across branches in the hierarchy (e.g. all 09:00–10:00 subsets). Here, we found interactive techniques to be effective, including the ability to switch hierarchy, to use brushing to highlight equivalent nodes, to generate raster and road maps and to have access to numerical details on-demand. The spatial filtering facilitated by the map views allows local detail to be represented without resorting to dense crowded road maps for the whole study area.

Various example treemaps have been compared with other layout methods in usability studies that aim to assess their performance as data exploration tools,[27,28] some of which report that treemaps perform poorly when compared with other hierarchical visualisation methods. The treemaps used here are very different to those usually tested – ours are characterised by fixed-sized leaves and spatial and temporal ordering and are employed in a complex task in which we are exploring a large multivariate spatial data set rather than finding items in a small hierarchy. Our informal analysis of interactive ordered treemaps as described and used here for variable selection in spatio-temporal visualisation is positive. We would be interested in assessing other space-filling visualisation methods for this purpose.

Technologies

The technological challenge of working with such a large data set was considerable and this had an impact on some of our design decisions. We used a PostgreSQL[30] database to store, maintain and query the data and generate output for visualisation. The PostGIS spatial extensions[31] enabled us to snap GPS points to their closest road segment. The high computational overhead associated with this operation forced us to use a randomly sampled 16 million point subset (38%). Initial experiments showed even small random samples appeared to be representative of the data set as a whole (we tried 1 million, 3 million and 16 million point samples). The main problem is that some subsets have sample sizes that are too small for meaningful summaries to be generated (as apparent in Figure 7D).

The large size of the data set necessitated the precomputation of all the numerical summaries used to generate graphics and for incorporation into the interactive design (this can be scripted and generated without intervention), reducing the flexibility of the approach some-

what. We chose to use the open technologies of SVG and Javascript to build our interactive system and found these to perform adequately, even though thousands of pregenerated numerical summaries need to be provided to the user on-demand. A significant limitation of the technologies used is that although they are based on well-documented standards, they are inconsistently implemented by browsers. They work well in Microsoft Internet Explorer under Windows with Adobe's SVG plugin and adequately in Safari on MacOS X, but more implementation work is required for consistent interpretation by more browsers.

Further work

The large size of our data set has had an impact on the design and functionality of our system, but we are looking to ways in which we can add more interactive filtering. Combining some of the relatively arbitrary subsets used here (e.g. vans and large vans, 10:00–11:00 and 11:00–12:00 or Angel and Farringdon) would not require a large computational overhead and could be achieved using the technologies and methods we employ – allowing the data set to be studied at different levels of granularity. Filtering that involves re-aggregating the original point data – such as filtering out subsets that contain few data points or considering trajectories with particular characteristics – is technically more challenging. Such queries are likely to be useful, and work to explore some of these possibilities is ongoing.

We are also looking at techniques that measure similarities and differences between subsets, rather than relying entirely on visual inspection. This might help alert the user to potentially interesting combinations and would draw this work closer to that of the geovisual analytics research agenda.[8]

We have used established theory, our own ideas and experience, and the published techniques and suggestions of others to develop these methods. In so doing we have developed our knowledge of the eCourier data set. We are engaging with other users of the eCourier and similar data sets to establish whether such techniques can help meet the goals of their specific exploratory analysis tasks.

Conclusion

This work was motivated by the desire to interpret and evaluate a large data set representing the characteristics of eCourier vehicle traffic through an open API.[9] The data set has the challenging properties of being large and multivariate with a dense spatial structure but likely to contain strong spatio-temporal patterns relating to traffic usage in London. We have designed, developed and reflected upon novel techniques for the visual exploration of this large and complex multivariate spatio-temporal data set that has posed substantial challenges in terms of generating meaningful aggregated summaries, providing access

to specific and selected detailed information and interactively linking these two processes to support exploration.

Our approach involves providing a rich overview of the data set through new treemap techniques that visualises thousands of variable-constrained subsets simultaneously in a single data-dense graphic in which appropriate and consistent ordering is used to facilitate the identification of patterns through space, time and attributes. Our interactive design allows us to select hundreds of subsets of interest from these rich graphics for further investigation. This is achieved through linked views through which we can compare the summaries of selected subsets with global traffic volume (using interactive treemaps), explore their variation across space (using raster maps) and consider the detailed distribution on the road network in localised areas (using road maps). Our enhancements to the ordering mechanisms used in treemaps extend their suitability to geovisualisation techniques. We acknowledge some of the widely cited weaknesses of treemaps and address some of these by using appropriate and consistent node ordering, hierarchy switching and interactive techniques. These support hierarchy depth changing, brushing to highlight equivalent nodes across the hierarchy, interactive linking and techniques for zooming, filtering and providing details on demand (Figure 7).

We have shown examples of how these techniques have enabled us to find structure and patterns in the data, some of which may be difficult to identify with alternative methods. The most important benefit of the treemap technique is the multifaceted overview of the entire data set that can be generated, such as those shown in Figure 4, due to the consistent spatial and temporal ordering, which effectively provide visual signatures of the traffic characteristics. Although the larger traffic volumes and slower speeds in the central area are expected patterns, some of the differences in the traffic composition are not so expected. For example, the central grid squares in Figure 4 (top) show similar signatures of high van traffic (upper left) at all times of the day, that motorbikes (lower left) and large motorbikes (lower right) show strong diurnal variation and that their use varies spatially.

The interactive design provides an interface that makes it possible to rapidly switch from an aggregated overview to a detailed interactive view in which the statistics of individual road segment can be retrieved and, in our example in Figure 7D, used to assess the appropriateness of the statistics. The detailed visual analysis in Blackfriars suggests that much of the high volume of 'van traffic' does not, in fact, represent through traffic. This type of finding suggests spatial bias that has implications for alternative uses of the data set.

We recommend further work in using treemaps and other information visualisation techniques for representing variable selection combinations that include time and space as more and larger spatio-temporal data sets become available through similar APIs to that provided and maintained by eCourier.[9] There is also scope for empirical cognitive studies that examine the effectiveness of combining different spatial, temporal and thematic layouts of treemap nodes in the same representation.

References

1 Keim D, Hao MC, Ladisch J, Hsu M, Dayal U. Pixel bar charts: a new technique for visualizing large multi-attribute data sets without aggregation. In: *Symposium on Information Visualisation 2001* (San Diego, CA), IEEE Computer Society: Silver Spring, MD, 2001; 113–122.

2 Chen J, MacEachren AM, Guo D. Supporting the process of exploring and interpreting spacetime multivariate patterns: the visual inquiry toolkit. *Cartography and Geographic Information Science* 2008; **35**: 33–50.

3 Keim DA. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 2002; **8**: 1–8.

4 MacEachren AM, Wachowicz M, Edsall R, Haug D, Masters R. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science* 1999; **13**: 311–334.

5 Gahegan M. Beyond tools: Visual support for the entire process of GIScience. In: Dykes J, MacEachren AM, Kraak M-J (Eds). *Exploring Geovisualization*, Elsevier Ltd: Amsterdam. 2005; 83–99.

6 Robinson AC, Chen J, Lengerich EJ, Meyer HG, MacEachren AM. Combining usability techniques to design geovisualization tools for epidemiology. *Cartography and Geographic Information Science* 2005; **32**: 243–255.

7 MacEachren AM, Kraak M-J. Research challenges in geovisualization. *Cartography and Geographic Information Science* 2001; **28**: 3–12.

8 Andrienko G, Andrienko N, Jankowski P, Keim D, Kraak M-J, MacEachren AM, Wrobel S. Geovisual analytics for spatial decision support: setting the research agenda. *International Journal of Geographical Information Science* 2007; **21**: 839–857.

9 eCourier. eCourier API [WWW document] http://api.ecourier.co.uk/ (accessed 12 June 2008).

10 eCourier. Ecourier News [WWW document] http://www.ecourier.co.uk/news.php (accessed 12 June 2008).

11 Department for Transport. DfT public service targets: technical note – PSA target 4 [PDF document] http://www.dft.gov.uk/pdf/about/howthedftworks/psa/spendingreview2004psatargets2 (accessed 12 June 2008).

12 Department for Transport. Tackling congestion on our roads [PDF document] http://www.dft.gov.uk/pdf/pgr/roads/roadcongestion/ (accessed 12 June 2008).

13 Goodchild MF. Citizens as sensors: the world of volunteered geography. *GeoJournal* 2007; **69**: 211–221.

14 Dykes J, Purves RS, Edwardes A, Wood J. Exploring volunteered geographic information to describe place: visualization of the 'Geograph British Isles' collection. In: *Proceedings of GIS Research UK* (Manchester, UK), 2008; 256–267.

15 Thomas JJ, Cook KA. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center: 2005; 190pp, ISBN is 0-7695-2323-4. http://nvac.pnl.gov/.

16 Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. *Symposium on Visual Languages* 1996 (Boulder, CO), IEEE Computer Society: Washington, DC, USA, 1996; 336–343.

17 Scheiderman B. Tree visualization with tree-maps: a 2D space-filling approach. *ACM Transactions on Graphics* 1992; **11**: 92–99.

18 LeBlanc J, Ward MO, Wittels N. Exploring *N*-dimensional databases. In: *Proceedings of the First Conference on Visualization '90*: 1990 (San Francisco, CA), IEEE Computer Society: Silver Spring, MD, 1990; 230–237.

19 Feiner K, Beshers C. Visualizing n-dimensional virtual worlds with n-vision. *SIGGRAPH Computer Graphics* 1990; **24**: 37–38.

20 Wood J, Dykes J. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics* 2008; **14** (6): in press.

21 Bruls M, Huizing K, Wijk JV. *Squarified Treemaps*. 2000 [PDF document] http://www.win.tue.nl /~vanwijk/stm.pdf. (accessed 12 June 2008).

22 Harrower M, Brewer CA. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal* 2003; **40**: 27–37.

23 Beshers C, Feiner S. AutoVisual: rule-based design of interactive multivariate visualizations. *IEEE Computer Graphics and Applications* 1993; **13**: 41–49.

24 Fabrikant SI. Visualizing region and scale in semantic spaces. In: *The 20th International Cartographic Conference*: 2001 (Beijing, China), 2001; 2522–2529.

25 Skupin A, Fabrikant SI. Spatialization methods: a cartographic research agenda for nongeographic information visualization. *Cartography and Geographic Information Science* 2003; **30**: 99–119.

26 Tobler W. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 1970; **46**: 234–240.

27 Andrews K, Kasanick JA. Comparative study of four hierarchy browsers using the hierarchical visualisation testing environment (HVTE). In: *Proceedings of the 11th International Conference Information Visualization*: 2007 (Zurich, Switzerland), IEEE Computer Society: Los Alamitos, CA, 2007; 81–86.

28 Cawthorn N, Moere AV. The effect of aesthetic on the usability of data visualization. In: *Proceedings of the 11th International Conference Information Visualization*: 2007 (Zurich, Switzerland), IEEE Computer Society: Los Alamitos, CA, 2007; 637–648.

29 Blanch R, Lecolinet E. Browsing zoomable treemaps: structure-aware multi-scale navigation techniques. *IEEE Transactions on Visualization and Computer Graphics* 2007; **13**: 1248–1253.

30 PostgreSQL [WWW document] http://www.postgresql.org/ (accessed 12 June 2008).

31 PostGIS [WWW document] http://www.postgis.org/ (accessed 12 June 2008).