U P T A P

ESRC
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

vizLib: DEVELOPING CAPACITY FOR EXPLORATORY ANALYSIS IN LOCAL GOVERNMENT VISUALIZATION OF LIBRARY USAGE DATA

Robert Radburn Research and Information Team, Leicestershire County Council and

Jason Dykes giCentre, City University London

Four key concepts underpin this project:

- *Libraries:* Leicestershire County Council (LCC) has 54 libraries serving over 600,000 residents, but the service is facing significant challenges from Internet bookshops, increasing budgetary pressures and changes in modern lifestyles.

- *Visualization:* involves the use of interactive maps and statistical graphics to identify structure in large heterogeneous data sets.

- *People:* the Government is encouraging local authorities to use locally collected data more effectively, and to base service decisions around customer behaviour.

- T*raining:* is required to equip researchers in local authorities with the skills, experience, confidence and connections needed to explore and analyse large holdings of geographic information.

Research questions and objectives

The Leicestershire Library Services (LLS) TALIS database of 450,000 lending records was visually explored to develop hypotheses around key research questions:

- How does performance vary across the 54 libraries in Leicestershire?
- In which areas are the best customers living (based upon usage of the library service?
- Can the area you live in contribute to predictions of usage?

The overall objectives of the fellowship were:

- to develop key skills in the visual analysis of geographic information;
- to apply this knowledge to explore large volumes of customer data collected by LLS;
- to share the knowledge and skills and embed them in local authority practice; and
- to develop links between local government and academia.

Approach

The data exploration involved using and developing visualization methods and implementing queryable graphics through which key patterns could be suggested, insights derived and hypotheses established. The skills to undertake this analysis and utilise such tools are not generally possessed in local authorities, but are becoming more accessible through high-level languages and toolkits such as *Processing* (Fry and Reas, 2007), *Prefuse* (Heer *et al.*, 2005) and *ProtoVis* (Heer and Bostock, 2009).

Fry (2008) proposes seven data visualization stages that form a template to visualize any data set. His model (Figure 1) draws attention to the interdependencies between the various stages of visual data analysis and the core skills required to employ visualization. We used the model to identify the necessary skills and competencies and to structure the visualization process. Fry contends that there is a danger that failing to consider data visualization as a pipeline may result in activities becoming ends in themselves rather than part of an analytical process.
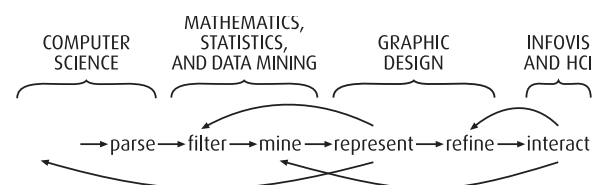

FIGURE 1. BEN FRY'S MODEL OF THE SEVEN STAGES OF VISUALISATION

represent — the visual models for the data

This short report focuses on stage five, where the following graphics were developed with *Processing* to represent the TALIS database, and to support the

exploration and research questions identified by LLS. Some of the most interesting patterns that help answer the three initial analytical questions are outlined below. These are a subset of the kinds of line of enquiry that resulted and continue to result from this analysis. Sophisticated and novel interactions and animated transitions were developed and used to relate the various views and queries in a highly dynamic manner.

User lending characteristics by library

RF plots: A matrix showing the different types of library users by combining the number of users into recency (column)/frequency (row) quintiles for each library. The most recent and frequent users, deemed by LLS to be the 'best users' of libraries, are located at the top right in Figure 2 (Novos, 2004).

In most cases, recency and frequency are related, hence the diagonal patterns. The extreme positions whereby users have high recency and frequency or low recency and frequency dominate. In Figure 2 we can see that the proportion of low recency/low frequency users (bottom left) in Melton Mowbray is far greater than that in the high recency/frequency category (top right). In Hinckley, a larger proportion of the least frequent users have visited more recently than those in Melton.
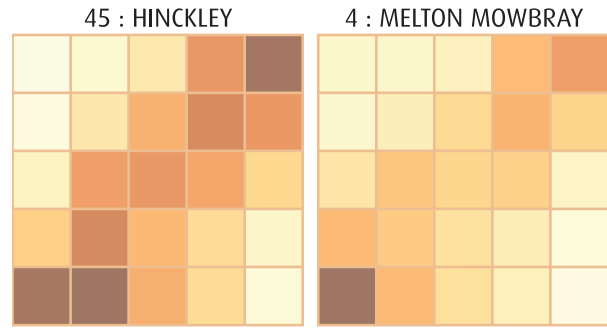

FIGURE 2. RF PLOTS FOR HINCKLEY AND MELTON MOWBRAY LIBRARIES

Creating RF Plots for all libraries (Figure 3) allows us to see variation within and between the different sites. Whilst the seven largest libraries dominate in terms of absolute numbers, scaling the colour scheme by a power function allows us to see variation amongst the smaller quantities.

The 'CHI comparison': Allows RF plots for libraries to be compared using a signed Chi statistic. The statistic compares an observed number with some expectation. The figure gives an indication of the degree of difference between observation and expectation. It accounts for the number of observations and is signed according to whether this is greater or less than the expectation. In Figure 4, the
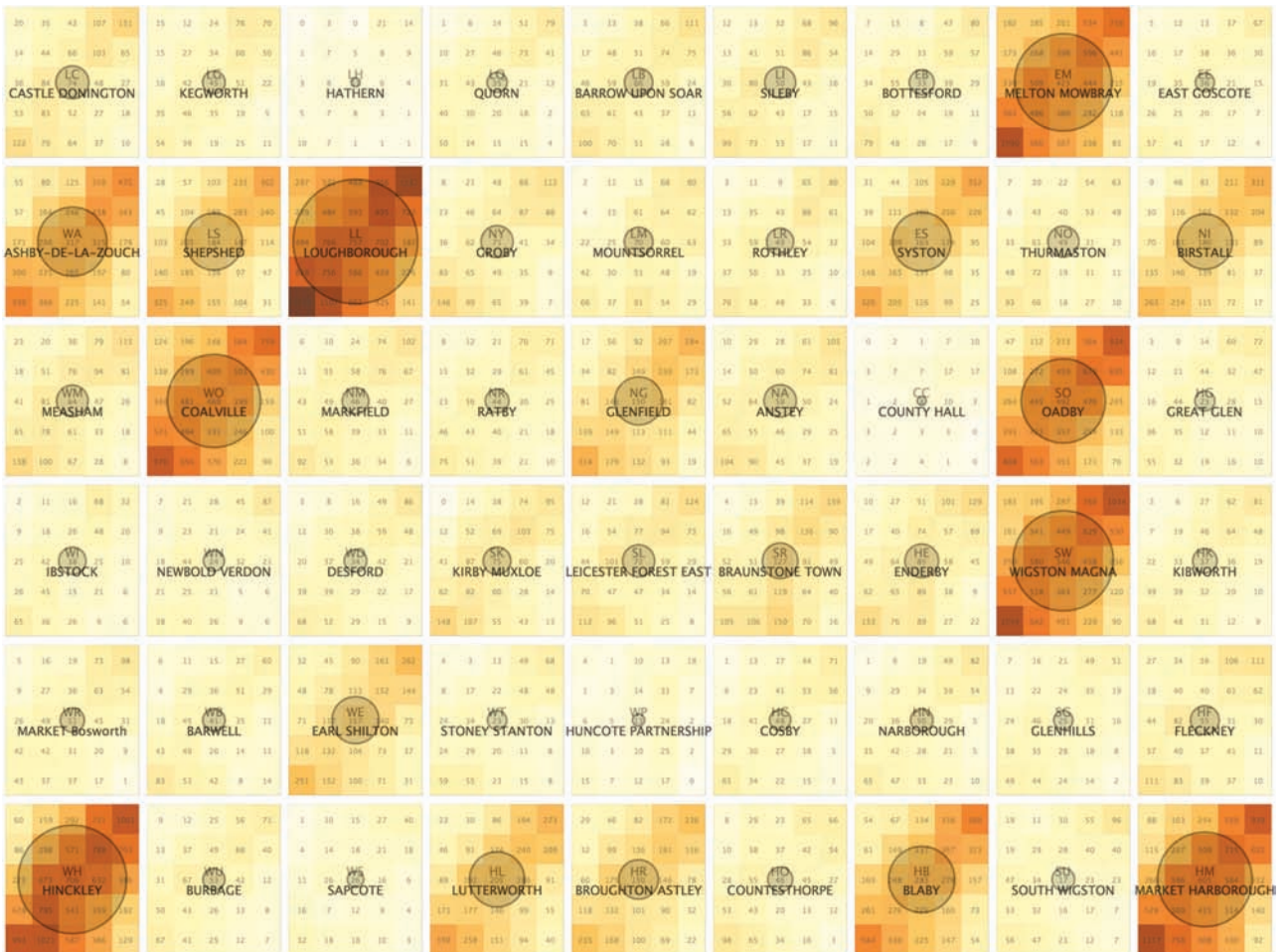

FIGURE 3. RF PLOTS MAPPED FOR 54 LEICESTERSHIRE LIBRARIES — FIXED SIZE SPATIAL TREEMAP SHOWING ABSOLUTE NUMBER OF USERS IN EACH OF 25 RF CATEGORIES FOR EACH LIBRARY WITH SEQUENTIAL 'YLORBR' SCHEME (DARKER SHADES HIGHER, LIGHTER LOWER). CIRCLE SIZE REPRESENTS NUMBER OF REGISTERED LIBRARY USERS

model upon which our expectation is based is one whereby each library can be expected to have the same proportion of users in each RF cell as the overall proportion for the county. Concurrently visualizing Chi statistics for all 54 libraries reveals higher than expected numbers of low recency users, shown in red in Figure 4, in some larger libraries (e.g. Melton, Coalville, Loughborough, Wigston) but not others (e.g. Hinckley).

User geography by library

Spider plots: Graphics representing the spatial relationships between libraries and customer locations; some of Market Harborough library's 'best users' have home addresses in villages that also have libraries as shown in Figure 5, where output areas (OAs) are coloured according to the most indicative super group from the Output Area Classifier (OAC — Vickers and Rees, 2007) to provide information about the nature of the origin areas.

Quartile plots: Concentric rings show the distances travelled by the closest 25%, 50% and 75% 'best users' (Figure 6), demonstrating that despite the long-legged spider plots, the majority of these citizens use their local library.

Standard ellipse/Weighted standard ellipse: Concentric circles do not account for any variation in the spatial

distribution of users. A standard ellipse of the 'best users'' home locations summarizes the point distribution, and reveals directional patterns in the origin data. In Figure 7, the road network and river seem to affect spatial usage patterns, because although Birstall and Thurmaston libraries are very close in proximity, there is very little geographic overlap between user home locations. This is useful information for LLS when deciding on opening hours, as the usage patterns suggest that the availability of one of these libraries cannot be regarded as necessary to cover the users from the neighbouring catchment.

User geography across Leicestershire

Spatial treemaps: We used the new 'spatial treemap' visualization technique for non-occluding space-filling layouts (Wood and Dykes, 2008). Symbols representing geographical units are sized according to population and arranged to reflect their locations and the hierarchical geography of the units in Leicestershire that they represent. Figure 8 shows Leicestershire output areas nested within wards and districts. Output areas are coloured according to a signed Chi statistic in which the number of registered library members is compared with that predicted by the county average. The cluster of blue-ish zones reveals a significantly lower number of members than expected registered in Thurmaston, leading to an examination of service provision
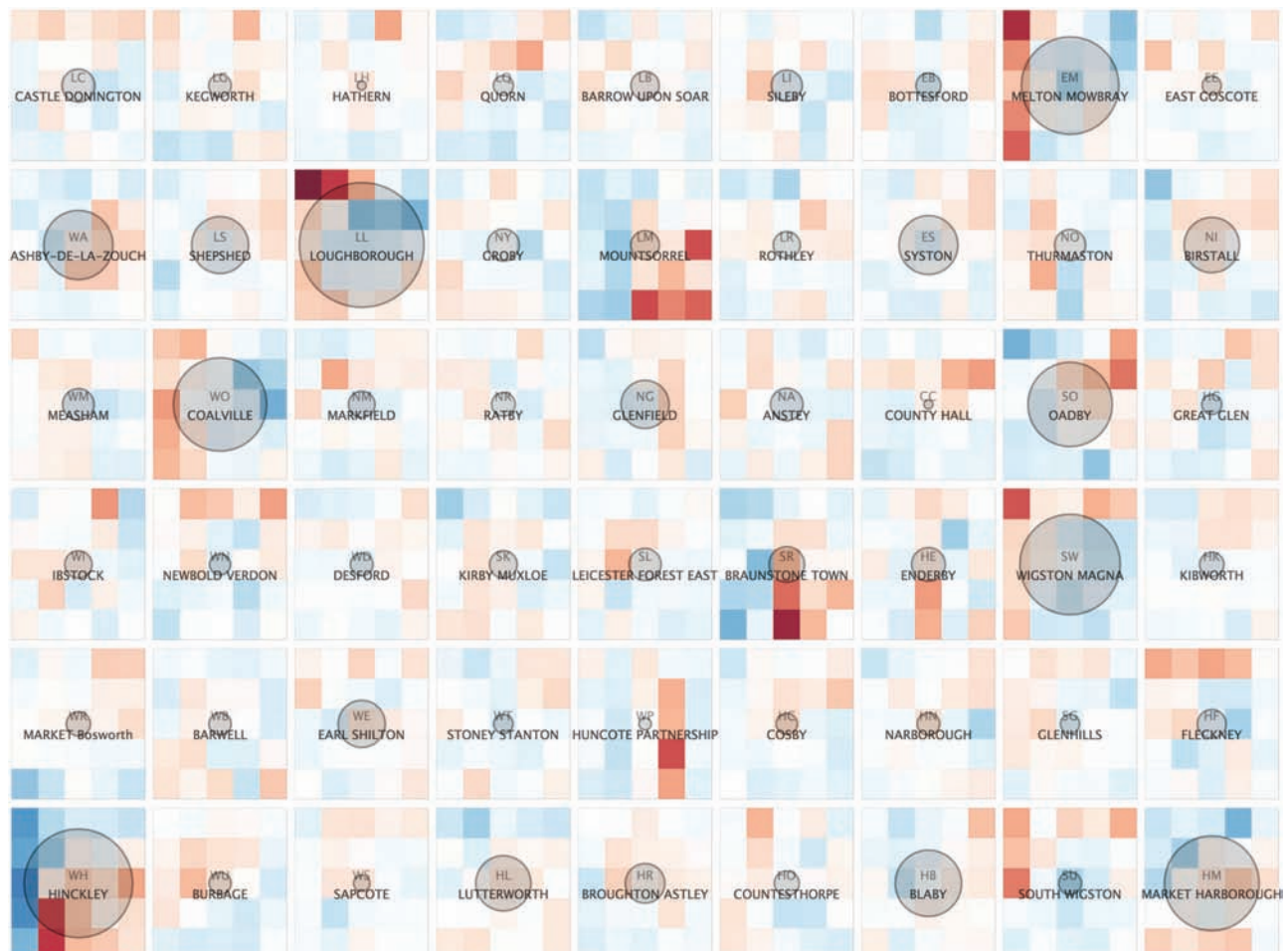

FIGURE 4. RF PLOTS MAPPED FOR 54 LEICESTERSHIRE LIBRARIES — FIXED SIZE SPATIAL TREEMAP SHOWING SIGNED-CHI STATISTICS WHEREBY NUMBER OF USERS IN EACH RF CATEGORY IS COMPARED WITH EXPECTATION BASED ON PROPORTIONS AT COUNTY LEVEL WITH DIVERGING 'RDBU' SCHEME (RED HIGHER THAN EXPECTED; BLUE LOWER). CIRCLE SIZE REPRESENTS NUMBER OF REGISTERED LIBRARY USERS
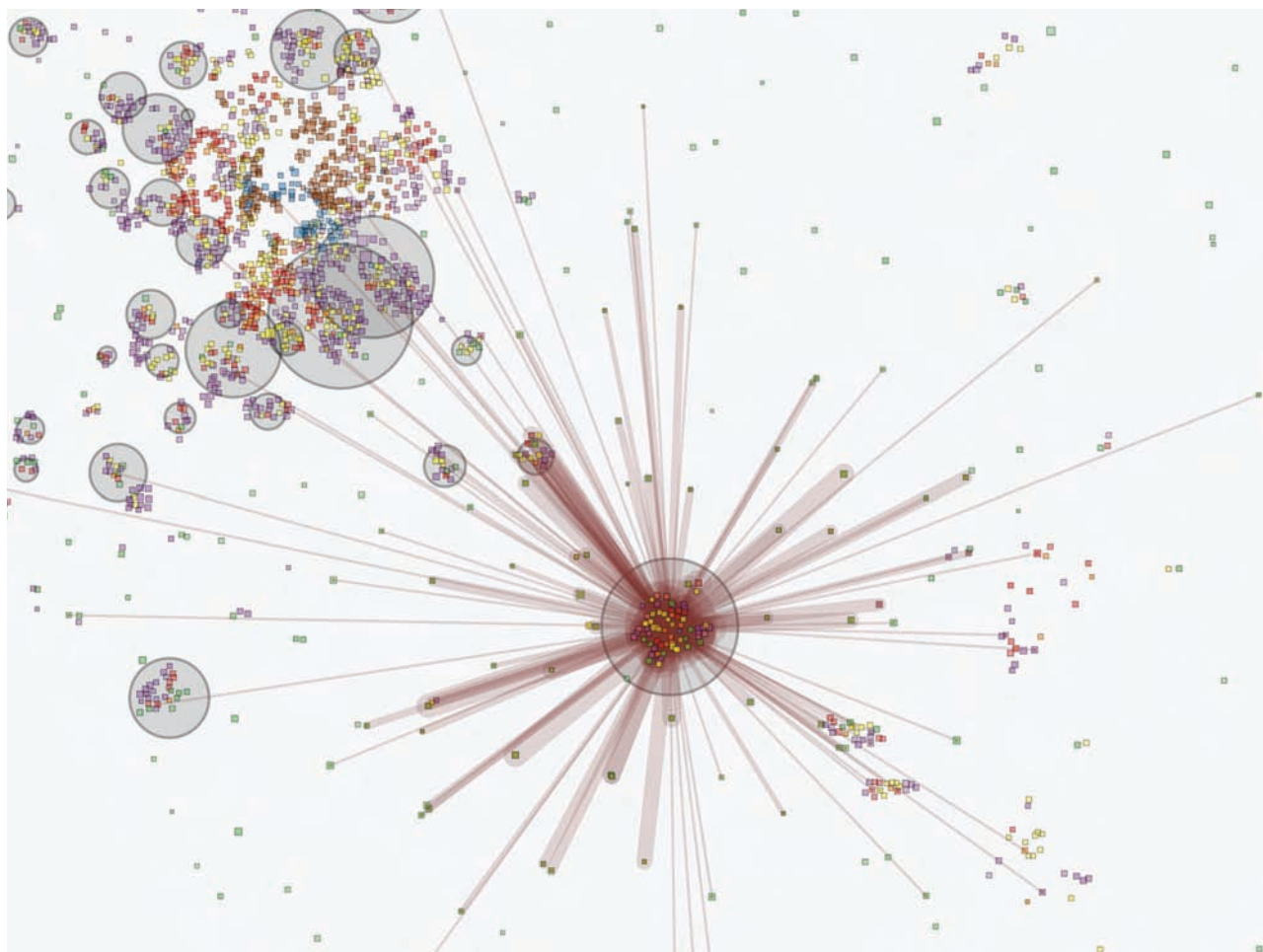
FIGURE 5. HOME LOCATIONS OF 'BEST USERS' OF MARKET HARBOROUGH LIBRARY BY OUTPUT AREA WITH COMPETING VILLAGE LIBRARY LOCATIONS. OUTPUT AREAS SIZED BY POPULATION, HIGHLIGHTED BY NUMBER OF RF 55 USERS AND SHADED BY OAC SUPERGROUP
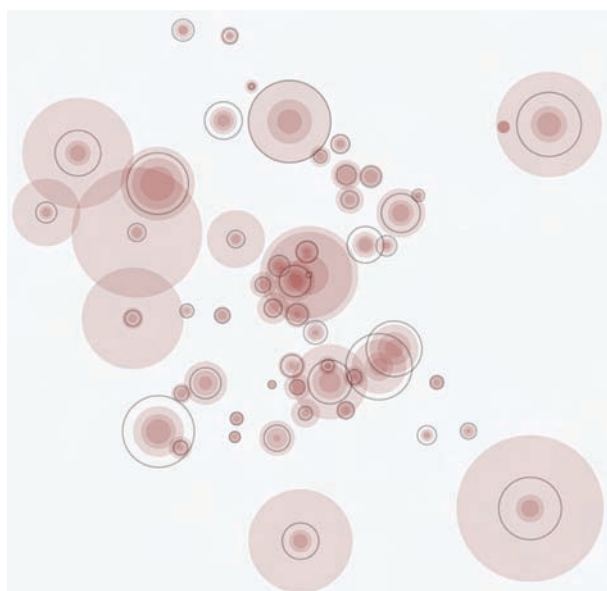

FIGURE 6. CATCHMENT AREAS FOR ALL 54 LIBRARIES WITH GEOGRAPHIC LOCATIONS. GREY LIBRARY SYMBOLS SIZED ACCORDING TO BORROWER POPULATION. CONCENTRIC RED SYMBOLS SHOW DISTANCES TRAVELLED BY 25%, 50% AND 75% OF 'BEST USERS'

in the area. The same is true of OAs in two other wards and a single large OA with a high population. Our spatial

treemaps morph smoothly into geographic layouts so that we can consider the geography to help explain these unexpected low levels of library membership.

Key findings
Processing allowed us to rapidly parse, filter, mine, represent, refine and interact with a large data set by developing novel dynamic graphics through which a huge number of broad and local, general and specific questions were asked of the data that could not have been considered previously. These queries have resulted in various hypotheses regarding population trends and processes in Leicestershire relating to the importance of transportation networks, geographical features and the spatial behaviour of customers using the service. This project has also generated considerable excitement within the organisation, and Leicestershire CC is looking to invest in further secondary data exploration and analysis using these techniques in other policy areas. One area under consideration is the daily flow of children to and from schools. The form of visualisation applied to the TALIS database has potential for helping local authorities understand this significant daily migration.

We should not underestimate the skills and processes required to follow Fry's framework in manipulating and making sense of large data sets from their acquisition
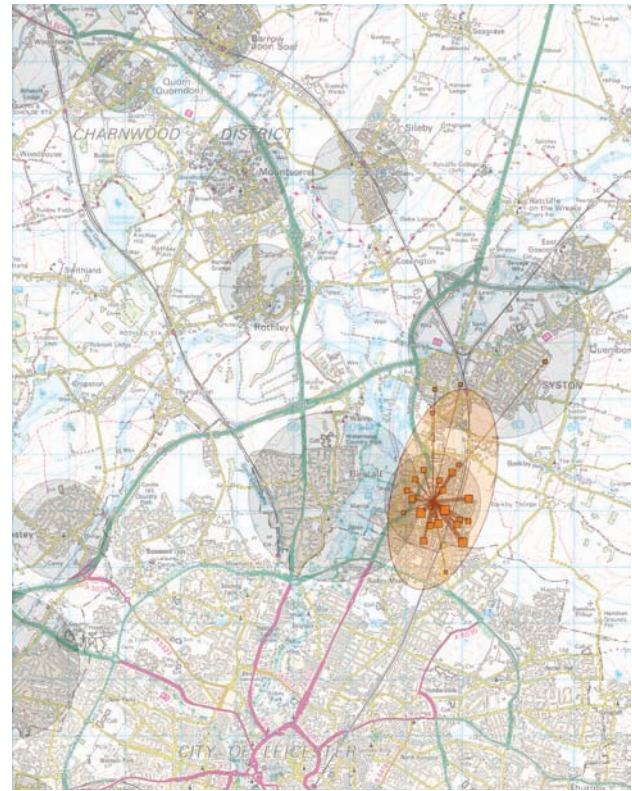
FIGURE 7. HOME LOCATIONS OF 'BEST USERS' FROM NEIGHBOURING BIRSTALL AND THURMASTON LIBRARIES WITH STANDARD ELLIPSE AND 1: 50,000 LAND RANGER *© Crown Copyright/database right 2009. An Ordnance Survey/EDINA supplied service.*

through visualization to decision-making. How do we decide what questions to ask? And how do we know what to do with the provisional and partial answers that are suggested as a result? Those developing policy face some new issues if they are to make the most of this kind of exploration of large data holdings. This requires significant attention if interactive visualization is to be used to its potential. Ultimately, if visualization is to have influence in organizations that hold large data sets, then a focus on an eighth stage of visualization may be important — **ACT**.

This project has led to an important local service beginning to fully understand its customers. Visualization techniques have allowed us to address the three research questions identified by the library service:

- the RF plots show that performance and user profiles vary across 54 libraries, reveal detail and allow comparison;
- the standard ellipse and spider maps give insights into where the 'best customers' live; and
- the spatial treemaps and catchment area maps allow us to consider usage patterns across the county concurrently and in relation to this, secondary information may enable us to explain and make predictions about usage by area.

Conclusions

The project has enabled the User Fellow to develop analytical capacity in his organization, where the methods and good practice are being shared with colleagues, and to collaborate on a paper with the giCentre using Leicestershire library data that won the Best Paper award at the GISRUK 2009 Conference. Through this visualization

work, the User Fellow was named 'LCC employee of the year in 2009 for innovation and improvement'.

We strongly support the concept of the User Fellowship, which has brought significant benefits to both individuals and institutions involved. If we as a society are to make use of the masses of information that we are currently recording about ourselves and **ACT** to improve the way in which society operates, skills such as those developed through this Fellowship will be essential. The kind of collaborative project between academic researchers and research analysts reported here can have a significant and beneficial impact in terms of linking research and practice and the learning that has been accomplished on both sides.

We use ColorBrewer schemes in all of our cartography.

References

Brewer, C. A. (2009) www.ColorBrewer2.org, [04/02/10].

Fry, B. (2004) *Computational Information Design*, Massachusetts Institute of Technology, Cambridge, MA.

Fry, B. (2008) *Visualizing Data*, O'Reily, Sebastopol, CA.

Fry, B. and Reas, C. (2007) *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press, Cambridge, MA, pp. 736.

Heer, J. and Bostock, M. (2009) Protovis: A Graphical Toolkit for Visualization, *IEEE Transactions on Visualization and Computer Graphics*, 15: 1121-1128.

Heer, J., Card, S. and Landay, J. (2005) Prefuse: A toolkit for interactive information visualization, *ACM Human Factors in Computing Systems*: 421-430.

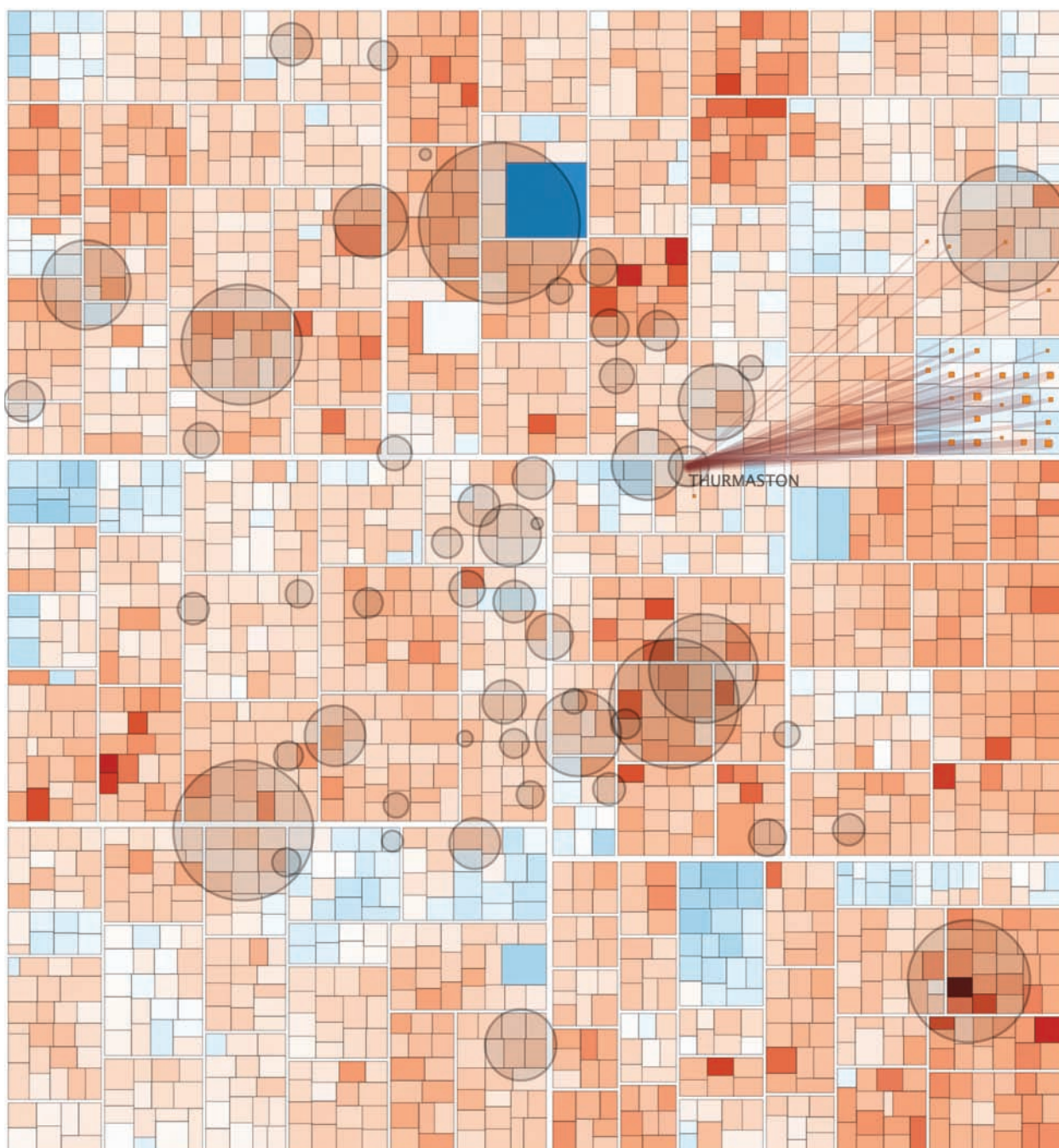Novos, J. (2004) Drilling down: Turning customer data into profits with a spreadsheet, booklocker.com.

FIGURE 8. SPATIAL TREEMAP OF LEICESTERSHIRE SHOWING LIBRARY MEMBERSHIP LEVELS FOR OAs WITH A SIGNED CHI STATISTIC

Vickers, D.W. and Rees, P.H. (2007) Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A,* 170(2): 379-404.

Wood, J., and Dykes, J. (2008) Spatially ordered treemaps, *IEEE Transactions on Visualization and Computer Graphics*, 14: 1348-1355.

Wood, J. , Dykes, J., Slingsby, A. and Radburn, R. (2009) Flow trees for exploring spatial trajectories, In *Proceedings of the GIS Research UK 17th Annual Conference GISRUK 2009*, University of Durham, Durham, pp. 229-234.

Contact Details of the Authors

Robert Radburn
Research and
Information Team,
Leicestershire County
Council, County Hall,
Glenfield LE3 8RA
Email: Robert.Radburn
@leics.gov.uk

Dr. Jason Dykes
giCentre,
Department of
Information Science,
City University London,
Northampton Square,
London ECIV OHB
Email: jad7@soi.city.ac.uk