Chapter 2

Visual approach to data analysis

Abstract
Unlike many other studies of cycling behaviour, this research relies on a dataset that was not collected solely for the purpose of doing social research. There is some uncertainty as to the nature and detail of research themes that might be addressed, but also as to the specific information, either already existing, external or derived, that might be used to answer research questions. Visual approaches to analysis are particularly suited to this more speculative analysis context. Designing interactive visual interfaces, patterns of cycling behaviour are quickly discovered and themes within the LCHS dataset explored. New hypotheses about behaviour are suggested as a result of this exploratory analysis, along with a more concrete set of research questions and analysis tasks. The use of visual data analysis methods therefore helps in moving to a point where more specific research questions are addressed and this is reflected in some of the later chapters. Within information visualization, this approach might be regarded as a *design study*: visual analysis software is designed to tackle an applied research problem. A pitfall particular to design studies, and to 'computational social science' research, is that, preoccupied with novelty in visual design, they fail to address problems within a target domain. This issue is perhaps sidestepped in this research by consciously designing and publishing analysis outputs within Transport Studies and closely involving colleagues at TfL, themselves specialists in transport policy, in the analysis.

2.1 Dataset and task uncertainty

This study aims to contribute new insights into research investigating how and why individuals cycle within cities. This literature was briefly introduced in Chapter 1 and more detailed reference is made to specific studies in the themed analysis sections. In most of this previous research, empirical datasets were collected for the purpose of studying specific aspects of cycling behaviour. For example, a researcher might be interested in the association between gender and claimed motivations and barriers to cycling (Heesch et al. 2012). S/he therefore designs a questionnaire and obtains a sample of research participants of appropriate structure and size such that these associations can be analysed within a statistical framework. Any limitations around the scope of analysis and strength of research findings are known in advance; so too, perhaps, are the specific analysis techniques likely to be used in deriving insights from the survey data.

The context under which this data-driven study is completed is necessarily different. The LCHS usage datasets were not necessarily generated for the purpose of doing social research. The customer database has a relatively sparse set of personal attribute information (see Chapter 3) and whilst the record of bikeshare journeys is spatially and temporally precise, important information such as journey purpose is not given. Further demographic information could be added by using external datasets and, through more detailed behavioural analysis, information on journey motivation might be derived. However, in studying LCHS cycling behaviour, this research focusses on a very particular population of cyclists with access to a very particular form of cycling. *Individual* customers' spatiotemporal usage behaviours have yet to be studied in detail within the existing bikeshare research (see Chapter 1). There is some uncertainty around whether a discernible structure might exist within the LCHS and whether the scheme is used sufficiently frequently by individuals for regular patterns of activity, or genuine *travel* behaviours, to be explored.

2.2 *Design study* method

Sedlmair et al. (2012) argue that visual approaches to data analysis are particularly suited to such research contexts: where the specific tasks necessary to achieve the stated research objectives are not obvious and the level of information required to engage with research questions not absolute. These applied, problem-driven research projects, Sedlmair et al.

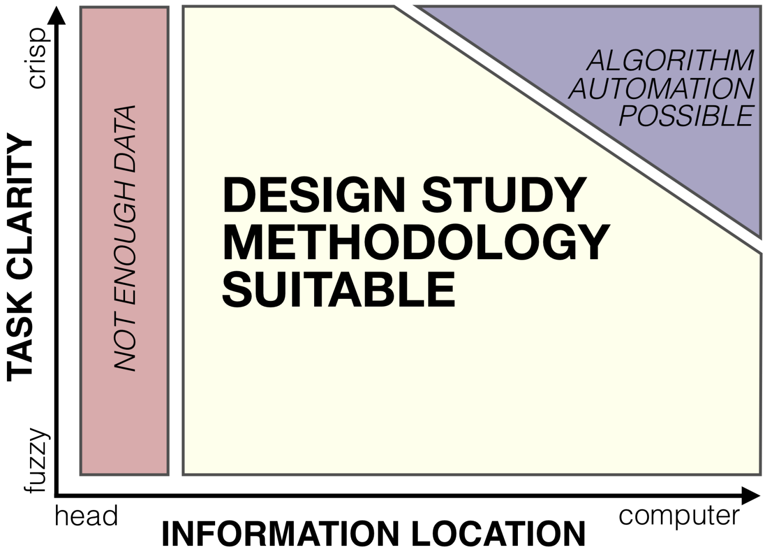

Figure 2.1: Research task and information space as appears in Sedlmair et al. (2012, p. 2433).

(2012) term them *design studies*, are generally exploratory in nature. They start with very specific and real-world problems, but with ambiguity around the research tasks and data models required to engage with these problems, their goal is to progress towards a point where the research tasks and the information used are more obvious and the analysis techniques increasingly automated. This is perhaps best captured in Figure 2.1, taken directly from Sedlmair et al.'s (2012) paper. Two conceptual axes are presented: task clarity and information location. Task clarity depicts how specific, stable and large the tasks necessary to answer a research problem are. Whilst task clarity can be both very precise or more nebulous, Sedlmair et al. (2012) suggest that domain problems are often not clearly defined. Information location characterises the extent to which data and contextual information required to carry out a set of analysis tasks are available to a researcher. In design studies, information location is never perfect. If it were, and if the tasks were well defined, visual analysis methods would not be needed. Instead, an existing set of algorithms might be automatically applied.

In this project, the three research questions are quite broad. Although the raw data might appear concrete (a set of customer and journey records), they alone are not particularly attribute rich and the extent to which modelled and external contextual data might also be used to answer these research questions is more ambiguous. As usage data are explored visually and patterns are discovered within the data, new hypotheses about behaviour are proposed, along with a more specific set of research questions and information tasks.

This movement from the more ambiguous research and information task to the more specific – a necessary contribution of a design study (Sedlmair et al. 2012) – is reflected through the analysis chapters.

The ambitions and research contributions of this research nevertheless do diverge in an important way from Sedlmair et al.'s (2012) conception of a design study:

> A design study is a project in which visualization researchers analyse a specific real-world problem faced by domain experts, design a visualization system that supports this problem, validate the design, and reflect about lessons learned *in order to refine visualization design guidelines.*
>
> (Sedlmair et al. 2012, p.2432; emphasis added)

Under this definition, 'visualization guidelines' perhaps become the object of analysis. By contrast, in this data-driven research it is the domain-specific insights and exposition of the LCHS dataset as a means to studying cycle behaviour, that is the main contribution. This privileging of application domain over computational novelty is justified with recourse to current critiques of the 'computational social sciences': that too often such data-driven studies are overly concerned with computational novelty and scalability across datasets and domains and, as a result, tend to lack analytical substance (Giles 2012, Watts 2013). This research does argue, with evidence, that a visual analysis approach is highly effective where new datasets are used for new purposes. It also demonstrates how visual techniques facilitate exploratory data analysis (Chapters 4 and 6) and support appropriate algorithm selection and development (Chapter 5). The main motivation, however, is not to abstract, refine and therefore contribute a unique set of visual design guidelines. Rather, existing guidelines are applied to a new application area and a case is made for their use in this and other application areas.

2.3 Problematising *design studies*

A number of factors apparently threaten the success of a design study. Many of these arguably apply to all research projects. Starting analysis before defining the research problem, before consulting with domain experts, before learning enough about a domain area and before sufficient data have been provided, are all early problems. So too is a design study that is not informed by literature. Implementing visual analysis techniques without recourse to visual perception theory and design principles is a problem particular to visualization research (Munzner 2008). Since in many cases a design study may be concerned with exploratory data analysis (Tukey 1977), the process of building analysis tools should be rapid. If *tool building* occupies a significant amount of researcher time, this serves as a brake on analysis, again a pitfall common in visualization research (Sedlmair et al. 2012). Also on the application of visual analysis techniques, Sedlmair et al. (2012) caution against studies that prioritise visualization novelty over domain-specific insights. Finally, Sedlmair et al. (2012) acknowledge that, when conducting a design study project there is rarely a distinct analysis or 'findings' phase. As is often the case in social science research, data insights are usually considered and eventually articulated at the writing phase (Sedlmair et al. 2012).

2.4 Visual analysis in this research

Sedlmair et al.'s (2012) warnings against design studies that are insufficiently grounded within a target domain and overly preoccupied with novel visualization technique, again seem particularly prescient given those same critiques have been levelled at the 'computational social sciences' (Miller 2010, Giles 2012, Watts 2013). In this research, a significant amount of time was spent early on with policy makers and database owners at TfL. The time was used to establish precisely what information could be made available, as well as any possible data protection issues associated with sharing information. A 'research problem' was identified through engaging with the Transport Studies discipline: attending and contributing at conferences, assimilating existing literature on cycling behaviour and also through discussions with policy specialists at TfL. Each analysis chapter starts with a précis of this relevant literature and how the analysis undertaken might contribute to that literature. To ensure that research themes and findings are policy-relevant, meetings with relevant contacts at TfL were also held in order to identify and characterise research needs. In terms of design, visual analysis techniques were implemented only after considering current research in information visualization and visual analytics (Chapter 3); and the programming environment used to develop visual analysis applications (*Processing*) has a set of libraries attached to it specifically designed for rapid prototyping. Finally, by discussing research findings with a wider group of policy makers at TfL, regularly presenting analysis techniques to academic and industry audiences and publishing discrete sets of analysis in relevant academic journals, research findings and their implications were routinely considered rather than simply deferring this sense-making activity to the final stages of the research project.

2.5 Moving forward

In this chapter, a visual approach to data analysis was briefly outlined. It was argued that visual approaches are particularly suited to this study – both to the dataset being used and to the research ambitions set out in Chapter 1. The application of techniques and approach to analysis in this context is a secondary contribution of this study. Despite the importance of approach, the work described here primarily aims to contribute to the Transport Studies domain. The link between analysis activities and domain-specific contributions is made clear in the analysis chapters. Each starts with a discussion of the research context and literature under which analysis activities are defined and to which contributions are made. Rather than including a high-level literature review chapter, then, relevant studies are incorporated throughout. The findings in Chapter 4, for example, support much of the existing literature on gender and urban cycle behaviour; in Chapter 6 new findings are offered on the subject of group cycling; and in Chapter 5 a novel analysis technique, given other recent analyses of large-scale OD transport datasets, is described. Rather than repeating lengthy descriptions of datasets and techniques in each of these analysis chapters, some time is spent in the following chapter discussing the datasets and derived variables used throughout the analysis, the main visual analysis application and some important underlying design principles.