geographical analysis

Geographical Analysis (2022) **0**, n/a

A Framework for Inserting Visually Supported Inferences into Geographical Analysis Workflow: Application to Road Safety Research

Roger Beecham[1], Robin Lovelace[2]

[1]Department of Geography, University of Leeds, Leeds, UK, [2]Institute for Transport Studies, University of Leeds, Leeds, UK

Road safety research is a data-rich field with large social impacts. Like in medical research, the ambition is to build knowledge around risk factors that can save lives. Unlike medical research, road safety research generates empirical findings from messy observational datasets. Records of road crashes contain numerous intersecting categorical variables, dominating patterns that are complicated by confounding and, when conditioning on data to make inferences net of this, observed effects that are subject to uncertainty due to diminishing sample sizes. We demonstrate how visual data analysis approaches can inject rigor into exploratory analysis of such datasets. A framework is presented whereby graphics are used to expose, model and evaluate spatial patterns in observational data, as well as protect against false discovery. Evidence for the framework is presented through an applied data analysis of national crash patterns recorded in STATS19, the main source of road crash information in Great Britain. Our framework moves beyond typical depictions of exploratory data analysis and transfers to complex data analysis decision spaces characteristic of modern geographical analysis.

Introduction

Road safety is a high priority area for those working across the transport planning, public health and engineering domains. It is a major reported barrier to uptake in active transport, discouraging affordable and healthy travel options and thereby contributing indirectly to the physical (in)activity epidemic (Winters et al., 2011; Sanders, 2015). Research findings generated from road safety datasets are often presented within a confirmatory data analysis framework. The aim is to test some prespecified hypotheses (Elvik et al., 2009) by quantifying the effect of a combination of selected explanatory variables on a clearly specified outcome variable (crash frequency or crash risk). This sort of confirmatory activity lends itself to research settings where there is certainty in input datasets, measurements and analytical techniques. In road safety

Correspondence: Roger Beecham, Department of Geography, University of Leeds, Leeds, UK
e-mail: r.j.beeecham@leeds.ac.uk

Submitted: January 13, 2022. Revised version accepted: May 25, 2022.

doi: 10.1111/gean.12338

research, however, such certainty is hard to achieve. When considering the factors that contribute to crash risk, relevant confounding effects and interactions may be difficult to specify in advance and may not be represented well by the available data. Even reasonably high-level comparisons of aggregated crash and injury risk are difficult to establish from road crash datasets. Reliable estimates of risk require appropriate denominators which may be scarcely available, exposure variables for pedestrian traffic are in particularly short supply; and where comparisons are to be made across data subsets of diminishing sample sizes, techniques must be deployed to guard against fitting models to noise and generating false discoveries. Separately, when analyzing geographic variation in crash risk, inferences are likely subject to change depending on the scale of spatial aggregation (Loidl, Traun, and Wallentin, 2016), but again there are no agreed-upon heuristics for deciding on an appropriate level of aggregation.

These sorts of difficulties are longstanding in geographical analysis, but the challenges of inference and out-of-sample replicability of geographic findings (Brunsdon, 2014; Brunsdon, 2016; Kedron et al., 2021; Wolf et al., 2021) gain special importance in the current era of data-driven geography (Kitchin, 2014; Miller and Goodchild, 2015; Singleton and Arribas-Bel, 2021). They have also attracted direct attention from applied data analysis more generally. In a recent discussion, Hullman and Gelman (2021a) reflect on several pitfalls in Confirmatory Data Analysis for applied statisticians, data journalists and data visualization researchers. Of particular interest is that of weak and overly simplistic hypotheses. Hullman and Gelman (2021a) and others (Heer, 2021) attribute this to an unhelpful binary thinking around Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA) and argue for an enlarged, model-based exploratory analysis workflow in which complex models are invoked and evaluated through heavy use of graphical methods.

Our paper adopts this rethinking by contributing a new framework for EDA targeted at modern geographical analysis. Within this framework, data graphics advance the model-based analysis in several ways: enabling researchers to narrow-in on candidate models and data generating processes; to characterize in detail variation or “surprise” in those models (e.g. Correll and Heer, 2017); and by offering flexible techniques for representing uncertainty and discouraging spurious discovery – for injecting rigor into EDA. We prove the effectiveness of this framework through an analysis of pedestrian crashes recorded in England between 2010 and 2019.

In the remaining sections of the paper we elaborate on the EDA-CDA dichotomy and challenges with applying this to geographical analysis and particularly analyses of STATS19. We then present a framework for analysis that relies on visualization to support model-based reasoning and evaluation and validate the framework through two separate analyses of vehicle–pedestrian crashes. In discussion we identify key features of the framework, and explain how they *transfer* to more widespread difficulties in claiming inferences from within modern geographical analysis (cf. Brunsdon, 2014; Brunsdon, 2016; Kedron et al., 2021; Wolf et al., 2021) – where research projects start with a high level of uncertainty, or “researcher degrees of freedom” (Kedron et al., 2021), and where standard inferential statistics cannot be easily applied, leading to “weak replicability” (Goodchild and Li, 2021).

Background

The EDA-CDA dichotomy

Concerns around the replicability of published scientific findings have over the last decade or so prompted much introspection around the way in which knowledge is produced in scientific research (Open Science Collaboration, 2015; Amrhein, Trafimow, and Greenland, 2019; Devezer et al., 2021; Szollosi and Donkin, 2021). To discourage data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that statistical findings should principally be claimed through out-of-sample hypothesis tests (Devezer et al., 2021). As an organizing framework for this, data analysis is separated into two discrete phases: EDA and CDA. EDA is used when encountering a dataset for the first time to identify its high-level properties and structure and to delimit the scope of further analysis. Typically this is through the heavy application of data graphics rather than statistical models. At this point the analysis stops, a set of hypotheses are recorded (preregistered) and a new dataset is collected and used in a CDA to evaluate whether the new data are consistent with the preregistered hypotheses.

While this "*EDA* → *preregistration* → *CDA*" framework may address issues of data fishing, it may not always lead to strong theory and knowledge development. Too heavy a focus on hypotheses amenable to testing may result in knowledge statements that are overly specific or that do not express sufficient detail or richness in outcomes (Szollosi and Donkin, 2021). In a recent discussion in the Harvard Data Science Review, Hullman and Gelman (2021b) make the case for a larger role, or a realignment, of "exploratory" approaches. They argue that model development should be intrinsic to EDA. Rather than simply displaying descriptive summaries of observed data, an EDA should support inferential thinking by encouraging analysts to consider, and subsequently model for, the processes that might have generated that data. A key aspect of exploratory analysis is then enabling detailed comparisons against those reference distributions. If the ambition of a CDA is to accept/reject pre-specified hypotheses, then EDA is concerned with "*the particularities of the discrepancies between model and data*" (Hullman and Gelman, 2021b) − or locating and characterizing *where* the data depart from the reference distribution. Data visualization is instrumental to this sort of pattern-finding activity. Hullman and Gelman (2021b) and others (Cook, Reid, and Tanaka, 2021; Heer, 2021; Hullman and Gelman, 2021a) envision data graphics supporting iterative comparison to models of increasing complexity and sophistication. This may happen either explicitly, for example in graphical inference where plots are compared to others generated under simulated data (Buja et al., 2009; Wickham et al., 2010; Morris, White, and Crowther, 2019), or indirectly by the way in which the plot is composed and read.

STATS19 and its analysis

The *EDA* → *preregistration* → *CDA* workflow is best suited to mature research settings where the theories and methods are well-developed and where prior studies and data exist that can be used to judge and evaluate observed effect sizes (McIntosh, 2017). In modern geographical analysis settings, where administrative and passively collected data are variously combined and repurposed (Singleton and Arribas-Bel, 2021), these certainties are difficult to achieve. STATS19 is one such dataset where the analysis space is broad and there is no stand-out approach to formulating research questions, selecting datasets and applying statistical techniques.

The name STATS19 refers to a form completed by police in Great Britain to record road crashes resulting in injury. Data are published back to 1979 although in this paper we focus only on pedestrian injuries for crashes dating from 2010. STATS19 is spatiotemporal and attribute-rich, with many variables and so numerous confounding factors that can be explored. The dataset contains three tables: information on `accidents`, or `crashes`, and the corresponding `vehicles` and `casualties` involved. These tables consist mainly of categorical variables whose category values vary in detail. The `crashes` data contains the date-time and location of crashes as well as detailed information on the context under which the crash took place – the road class, type, speed limit and condition; details around the junction and road infrastructure involved; around the weather conditions and degree of lightness. The `vehicles` table contains data on the type, age of vehicle, the manoeuvre being made, but also detail on the driver – age, sex, geodeomgraphics (the Indices of Multiple Deprivation [IMD] class of their home location) and journey purpose. As well as identifying the severity of injury (slight, serious, fatal), the `casualties` table contains data on the age, sex, geodeomgraphics of the casualty, whether they were a driver, passenger or pedestrian.

Two focuses for STATS19 analysis are in establishing determinants of road crashes (e.g. Aldred et al., 2018; Kondo et al., 2018; Sarkar, Webster, and Kumari, 2018) and developing indicators for comparison and monitoring of risk. For the former, work is often incremental, with claimed findings around determinants validated through comparison across studies. These studies, which tend to be presented more as CDAs (though rarely using preregistration), require complex and detailed thinking around outcome and explanatory variables – their inclusion, numerical representation and interactions – and therefore some sophistication in both data and technique. Aldred et al. (2018), for example, attempt to identify the effect of road infrastructure on cycling injury risk in London. Crucially, the authors wish to estimate these effects net of exposure over the London road network – the amount of cyclist and motorized traffic. This is in order to separate out the safety-in-numbers effect conferred from greater cycling activity attaching to particular infrastructure and parts of the road network. Despite its importance to estimating risk, Aldred et al. (2018) note that attempts to quantify exposure at a network level are rare in the road analysis literature on cycling, and that findings from studies that have been published sometimes conflict, demonstrating the difficulty of generating empirical knowledge that replicates from such observational analyses.

Work around road crash monitoring has attempted to quantify and compare crash risk between different travel modes (Beck, Dellinger, and O'Neil, 2007), spatial and administrative reporting areas (Eksler, Lassarre, and Thomas, 2008; Jones et al., 2008; Hindle, Hindle, and Souli, 2009; Eksler, 2010; Boulieri et al., 2017; Boss, Nelson, and Winters, 2018), spatially varying demographic context (Lovelace, Roberts, and Kellar, 2016) or by some other demographic irrespective of geography (Scholes et al., 2018). This sort of activity is very policy-relevant, identifying thematic areas, geographic and other detailed context (Boss, Nelson, and Winters, 2018) to which funding, infrastructure and further research effort might be targeted. Such risk monitoring might therefore be presented as more exploratory in nature. However, different from the caricature of EDA, detailed thinking around dataset and technique is again required to address two issues. Firstly, how to represent exposure or the "population at risk." The number of registered vehicles, estimated miles traveled and population size and density aggregated over some areal unit have been variously used. Second is around the uncertainty and potential for false discovery likely when comparing across areal units or data subsets that are based on low numbers of observed crashes.

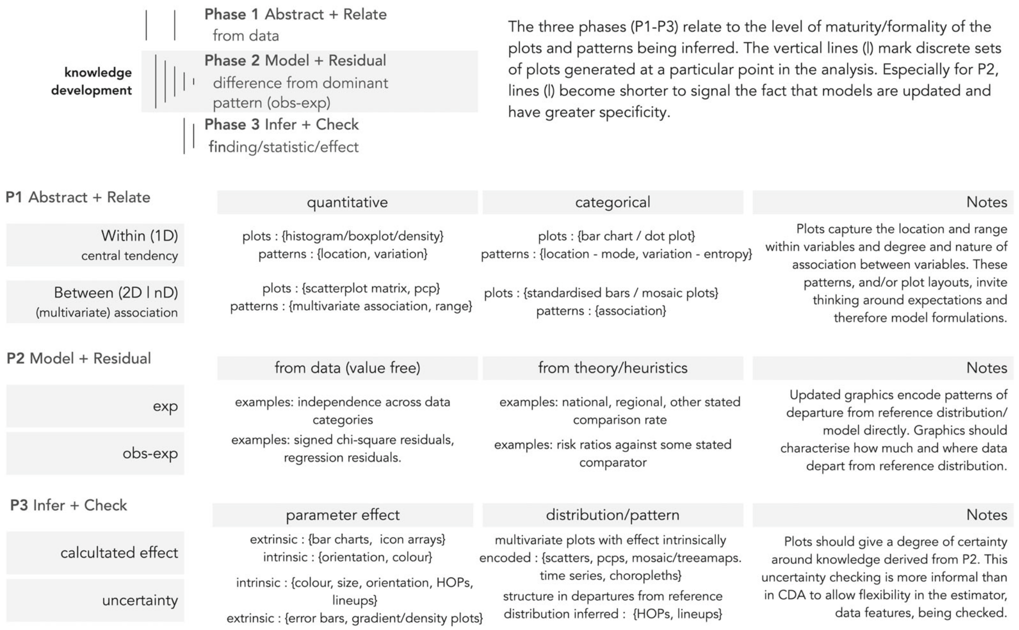
There is, then, not a single agreed approach to the analysis of STATS19 data. Many of the detailed categorical variables in the dataset remain underexplored, and so may be variously combined and reconfigured to address an undetermined set of research needs. Whilst it may be possible to specify in advance some high-level hypotheses to be tested formally through a CDA, the analysis possibilities are wide-ranging. As Hullman and Gelman (2021b) states, in such circumstances "[imprecision] *may in fact be better for encouraging a modeling-oriented mindset,* *where the goal is to learn about what assumptions describe the data better or worse, rather than to pose more dichotomous questions*." It is this positioning that motivates our framework and informs the application of visual analysis methods for generating inferences presented in the paper.

Framework

Our proposed framework attempts to synthesize the process of generating detailed knowledge from complex observational datasets, presented in such a way as to inform decision-making. The framework distinguishes between three different phases of analysis (*Phase1-Phase3* – Fig. 1), which generally relate to the level of maturity and formality of the plots and patterns being inferred.

Passing through these phases might not always result in the same outcome. On some occasions the upshot might be to collect more data for fuller analysis or to delineate some formal CDA-type activity; on others findings from one of the phases may be sufficient to be reported, meaning the focus shifts to the quantification and evaluation of "results." In cases where research outputs may inform real decisions, then reasonable steps to ensure robustness of findings – to check against false discovery – are necessary.

Although Fig. 1 suggests three distinct phases, in reality there is overlap and interaction between each of these. We demonstrate this through our implemented data analysis, where we spend some time oscillating between *Phase1* and *Phase2* (Section 5.1) and *Phase2* and *Phase3* (Section 5.2). *Phase2* is the key knowledge development phase, where data graphics are used to explore and evaluate models updated with increasing specificity. As such, we expect and



The three phases (P1-P3) relate to the level of maturity/formality of the plots and patterns being inferred. The vertical lines (|) mark discrete sets of plots generated at a particular point in the analysis. Especially for P2, lines (|) become shorter to signal the fact that models are updated and have greater specificity.

P1 Abstract + Relate	quantitative	categorical	Notes

Within (1D)
central tendency
	plots : {histogram/boxplot/density}
	patterns : {location, variation}
		plots : {bar chart / dot plot}
		patterns : {location - mode, variation - entropy}
			Plots capture the location and range within variables and degree and nature of association between variables. These patterns, and/or plot layouts, invite thinking around expectations and therefore model formulations.

Between (2D | nD)
(multivariate) association
	plots : {scatterplot matrix, pcp}
	patterns : {multivariate association, range}
		plots : {standardised bars / mosaic plots}
		patterns : {association}	

P2 Model + Residual	from data (value free)	from theory/heuristics	Notes
exp	examples: independence across data categories	examples: national, regional, other stated comparison rate	Updated graphics encode patterns of departure from reference distribution/ model directly. Graphics should characterise how much and where data depart from reference distribution.
obs-exp	examples: signed chi-square residuals, regression residuals.	examples: risk ratios against some stated comparator	

P3 Infer + Check	parameter effect	distribution/pattern	Notes

calcultated effect	extrinsic : {bar charts, icon arrays}
	intrinsic : {orientation, colour}
		multivariate plots with effect intrinsically encoded : {scatters, pcps, mosaic/treemaps. time series, choropleths}	Plots should give a degree of certainty around knowledge derived from P2. This uncertainty checking is more informal than in CDA to allow flexibility in the estimator, data features, being checked.

uncertainty	intrinsic : {colour, size, orientation, HOPs, lineups}
	extrinsic : {error bars, gradient/density plots}
		structure in departures from reference distribution inferred : {HOPs, lineups}	

Figure 1. The three phases of the framework for exploratory visual data analysis. The graphical techniques listed under P2 and P3 are documented with explanatory code examples in the paper's accompanying code repository (Beecham, 2022).

encourage iterative data analysis steps during this phase, with each iteration generating new insights into patterns revealed by the visual and descriptive analysis.

The framework is targeted at supporting model-building and knowledge development. Not covered are the more individual exploratory analysis protocols around processing geospatial data, described systematically in Graser (2021).

P1. Abstract + Relate

Phase1 (P1) matches most closely with the descriptions of EDA that appear in many textbooks on data analysis, usually invoking Tukey (1977). During this phase, datasets are described and organized according to their levels of measurement (Stevens, 1946). This informs the types of summary statistics and graphics that may be appropriate and how data might be aggregated. From here, graphics are generated that establish patterns within variables – their location and range – and relationships between variables. Fig. 1 provides a list of candidate plots and the likely patterns inferred depending on data type. These patterns and relationships should provoke thinking and help to inform expectations (null hypotheses) against what is likely to be interesting (alternative hypotheses). Plots might be further conditioned for comparison to explore and identify confounders. This might happen during the first pass of *Phase1* or on a second or third pass, after having evaluated some explicit model formulation at *Phase2*.

P2. Model + Residual

The model-based thinking initiated during *Phase1* is encoded explicitly at *Phase2* (P2). Initially this may be through a reasonably high-level model formulation derived from the data. Alternatively this model might be based on theory and heuristics. Data graphics are then designed to encode difference, or surprise (Correll and Heer, 2017), against this model and the nature of unexpected difference is used to extend and update the model specification, denoted in Fig. 1 by the narrowing of the parallel vertical lines. Different from a CDA where the analytic ambition is limited to model testing, the graphics generated at *Phase2* should provide rich detail around where in the distribution and by how much the observed data depart from the model. It is at this stage of close iteration between model and graphics that knowledge development occurs.

P3. Infer + Check

Phase3 (P3) introduces dedicated graphical techniques for checking inferences. In our implemented data analysis, we follow Hullman and Gelman's (2021a) suggestions and embed nonparametric bootstrapping within data plots to invite analysts to consider uncertainty implicitly. Fig. 1 enumerates some of these, for example hypothetical outcome plots (Hullman, Resnick, and Adar, 2015; Kale et al., 2019), ensemble displays (Quinan and Meyer, 2016; Liu et al., 2017) and graphical inference though line-ups (Buja et al., 2009; Wickham et al., 2010). In Section 5.2 such techniques are deployed when ranking local authority reporting areas according to crash rates and making inferences from observed geographic patterns in crash rates. Additionally the full set of uncertainty representation approaches listed in Fig. 1 is demonstrated through an analysis of injury severity in the paper's accompanying code repository (Beecham, 2022).

P3 is an important update to EDA approaches and a distinctive contribution of our framework. When analyzing model residuals in graphics it is often easy to select and isolate only the patterns and effects that appear most visually salient (Hullman and Gelman, 2021a). There is no guarantee that perceived effects are not artifacts of the way in which the plot is constructed and decoded; there may be a mismatch between effects that are visually salient and those that are statistically salient. In a CDA, expectations of false discovery are specified in advance and mitigating procedures baked-in to the research design. For the observational data analysis settings to which our framework is targeted, the inferences being made may not be fully expressed in the models. This is certainly true in geographical analysis, where complex locational, adjacency, distance and categorical relations may be analyzed concurrently in a way that is very difficult to encapsulate in a single model specification, and where issues related to the scale of aggregation, presence of spatial dependence and nonstationarity complicate the inferences that can be drawn (Kedron et al., 2021). In these instances we recommend that computational procedures and visual methods be used in tandem to promote graphical representations, and subsequent inferences, that are robust and sensitive to uncertainty.

Data-driven versus theory-driven analysis

While our framework aims to cut through the EDA-CDA dichotomy, the distinction between data-driven and theory driven analysis, articulated with respect to observational analysis in Elragal and Klischewski (2017), is useful for understanding the different decisions that take place through *P1–P3*. *Phase1*, for example, is initially data-driven. Datasets are described abstractly according to their measurement level and from here corresponding graphics and assumptions around distributions and relationships between variables are generally "value-free." However, this phase also demands some prior knowledge or heuristics – in our Data Analysis 2, an awareness of the competing problems of using residential and workplace population denominators for quantifying road crash risk. The need to apply theory, or existing knowledge, remains as more complex models are proposed and evaluated at *P2*. Researcher expertise gains particular importance at *P3*. Although at this phase techniques for estimating uncertainty are employed, in geographical analysis where many features are explored and many comparisons made concurrently, it is not obvious what level of uncertainty is acceptable when checking patterns against estimated uncertainty (Goodchild and Li, 2021). This applies in our documented analysis, especially Data Analysis 2. At this point, then, it is necessary for the researcher to exercise some judgement based on prior domain knowledge.

Data

The STATS19 dataset used to demonstrate and validate our framework consists of 10 years' worth of vehicle-pedestrian crash data – crashes that resulted in pedestrians being injured and those injuries being recorded by the police. Relative frequencies with which pedestrian crashes occur are examined at key reporting area units – Local Authority Districts (LADs). Detail around the characteristics of the vehicles and pedestrians involved in these crashes are compared. For this, the three distinct `Crashes`, `Casualties,` and `Vehicles` tables were linked, with some simplifications and assumptions made when associating single vehicles (drivers) with single pedestrian casualties. This process is detailed in the paper's accompanying code repository (Beecham, 2022).

A focus for Section 5.1 of our analysis is the geodemographic characteristics of crash locations, and of the pedestrians and vehicles involved in each crash. Collected in the STATS19 dataset is the IMD decile (Noble et al., 2019) of the small area neighborhood in which casualties and drivers live. Often IMD is reported at the quintile level, and we use quintiles rather than deciles in our analysis. IMD measures deprivation for LSOAs in England and so when joining this on the crashes dataset we analyze only crashes taking place in England.

In order to estimate crash frequencies between LADs it was necessary to standardize these frequencies in some way. The convention is to use population denominators for comparing relative differences in pedestrian casualties. Data from the Office for National Statistics' mid-year population estimates (ONS, 2021) was used, but initial analysis showed systematic biases in this choice of denominator: it under-estimated "exposure" in denser urban centers. Additionally, we collected data from 2011 Census describing the "workplace population" of each LAD and merged the two datasets in order to generate an updated estimate of population exposure. It should be noted that this is still a flawed indicator of exposure in that there may be systematic unmeasured differences between LADs in the relative proportions of those populations that are road users.

Crash data were collected using the `stats19` R package (Lovelace et al., 2019). Full materials and documentation can be found at the paper's accompanying code repository (Beecham, 2022).

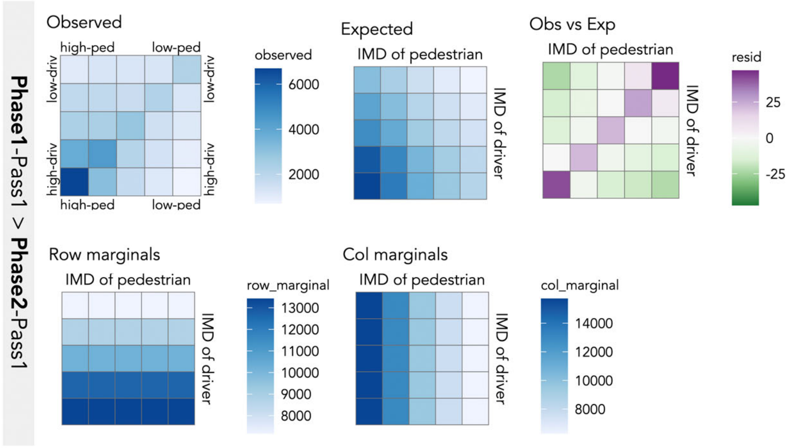
Application

Detailed in this section are two data analyses: the first explores inequalities in the demographics of those involved in pedestrian crashes (Section 5.1); and the second quantifies and compares under uncertainty pedestrian crash risk between geographic reporting areas (Section 5.2). The two analyses cover different sections of our framework and for each we identify how specific aspects of the framework, and accompanying graphics, are used to support model-based reasoning.

Data analysis 1: Inequalities in pedestrian-driver-location characteristics

Inequalities and road casualties is an important theme in road safety analysis. Research suggests those living in more deprived neighborhoods are at elevated risk of road crash injury than those living in less-deprived areas (Feleke et al., 2018; O'Toole and Christie, 2018; Tortosa et al., 2021). A follow-up question, especially relevant for crashes involving pedestrians, is around the characteristics of those involved in crashes. To what extent do drivers share demographic characteristics with the pedestrians they crash into, and does this vary by the location in which crashes take place? This theme has yet to be investigated using observational crash data, to the best of our knowledge. The analysis presented in this section investigates vehicle–pedestrian crashes in STATS19 between 2010 and 2019, for which vehicles could be linked to pedestrians and where the IMD class of the pedestrian, driver and crash location is recorded.

We wish to first profile how the characteristics of pedestrians involved in crashes, the drivers crashing into them and the locations in which crashes occur co-varies. To do this we *abstract* (*Phase1-Pass1*) over the dataset and initially consider the variables from which this profile is to be drawn: five IMD classes from high-to-low (IMD quintile 1–5) for pedestrians, drivers, and crash locations. We generate 5 × 5 contingency tables of the joint frequency of each permutation of driver-pedestrian IMD group co-occurring, colored by frequency and with cells ordered left-to-right by IMD class of pedestrian and bottom-to-top by IMD class of driver, as depicted in Fig. 2. This arrangement encourages linearity in the association to be emphasized. The darker blues in the diagonals of the top left graphic of Fig. 2 demonstrate such an association between driver–pedestrian geodemographics exists: drivers and passengers living in similar types of neighborhoods are involved in crashes with one another with greater frequency than those living in different types of neighborhoods. That color concentrates in the bottom left is also to be expected. A large share of high-deprivation neighborhoods are located in urban areas (DefRA, 2018), where pedestrian crashes are more likely to occur.



Obs: Larger crash counts for higher deprivation ped-drivers. Clear positive association between crash counts and IMD class of ped-driver.

Exp: Net of differences in absolute counts by IMD class, crash counts distribute independently of ped-driver IMD class.

Stat: Signed chi-square residuals give saliency to relative and absolute numbers in obs-exp.

Figure 2. Annotated analysis of pedestrian-driver geodemographics. [Color figure can be viewed at wileyonlinelibrary.com].

A consequence of the heavy concentration of crash counts, and thus color, in the high-deprivation cells is that it is difficult to gauge variation and the strength of association between IMD class of driver and pedestrian in the lower deprivation cells. As a follow-up (*Phase2-Pass1*) we generate expected frequencies for each position in the contingency table on the assumption that crash frequencies distribute independently of the IMD class of the pedestrian-driver involved. The second row of graphics in Fig. 2 demonstrates how expectation is spread in the contingency table based on absolute numbers within rows and columns. Expected values for each cell (E_i) can then be derived from multiplying across these column and row marginals, standardized by the total number of observations:

$$E_i = \frac{C_i \times R_i}{GT}. \tag{1}$$

So for an observed cell in the contingency table (O_i), C_i is the column total of that cell; R_i is the row total of that cell; and GT is the grand total. By explicitly encoding these expected cell frequencies (Fig. 2), the assumptions in this model of independence are further clarified. We expect larger counts to appear in the high-deprivation cells, but that these counts do not co-vary depending on the IMD class of pedestrian and driver.

Signed chi-score residuals (Visalingam, 1981) are used to compare observed ($O_i ... O_n$) and expected values ($E_i ... E_n$) for each cell of the contingency table:

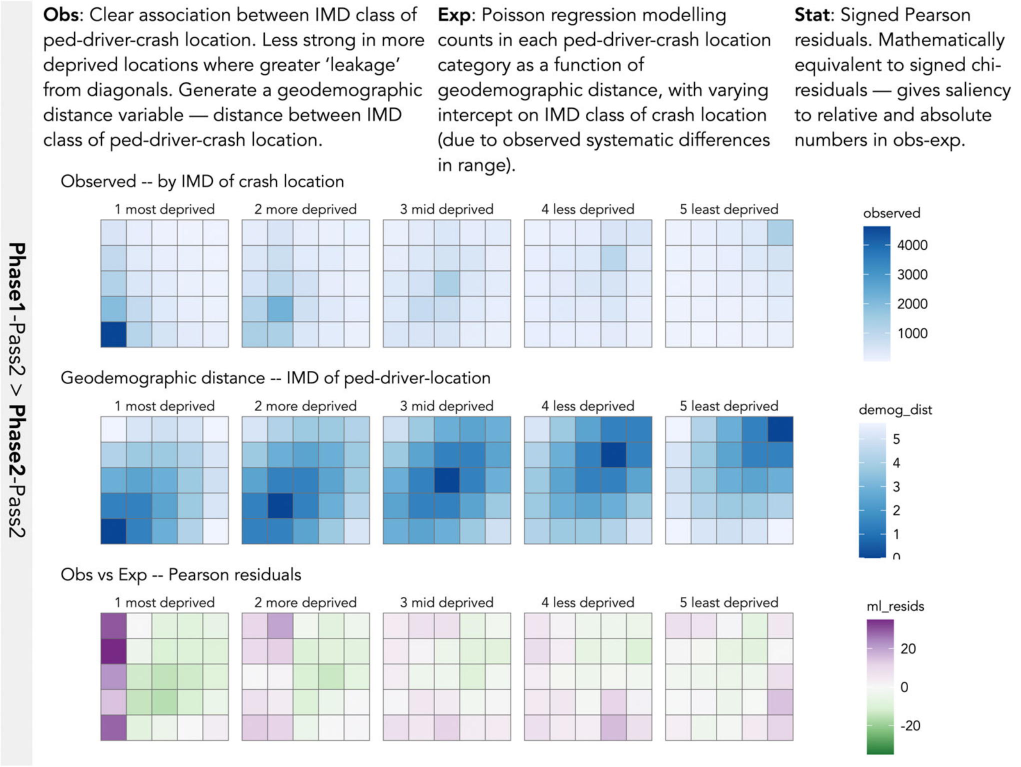
$$\chi_i = \frac{O_i - E_i}{\sqrt{E_i}}. \tag{2}$$

The way in which the differences (residuals) between observed and expected values are standardized in the denominator is important. If the denominator was simply the raw expected value, the residuals would express the proportional difference between each observation and its expected value. The denominator is instead transformed using the square root ($\sqrt{E_i}$), which has the effect of inflating smaller expected values and squashing larger expected values in the denominator, thereby giving greater salience to differences that are large in both relative and absolute number. These sorts of numeric properties are useful in road safety research, especially when data are subset and so are of smaller sample size. Signed chi-scores are mapped to a diverging color scale in the top row of Fig. 2: purple for residuals that are positive (cell counts are greater than expected), green for residuals that are negative (cell counts are less than expected).

The observed-versus-expected plot highlights that the largest positive residuals are in and around the diagonals and the largest negative residuals are those furthest from the diagonals: we see higher crash frequencies between drivers and pedestrians living in the same or similar IMD quintiles and fewer between those in different quintiles than would be expected given no association between pedestrian-driver geodemographics. That the bottom left cell – high-deprivation-driver + high-deprivation-pedestrian – is dark purple can be understood when remembering that signed chi-scores emphasize effect sizes that are large in absolute as well as relative terms. Not only is there an association between the geodemographics of drivers and casualties, but larger injury counts are recorded in locations containing the highest deprivation and so residuals here are large. The largest positive residuals (the darkest purple) are nevertheless recorded in the top right of the matrix – and this is more surprising. Against an expectation of no association between the geodemographic characteristics of drivers and pedestrians, there is a particularly high number of crashes between drivers and pedestrians living in neighborhoods containing the lowest deprivation. An alternative phrasing: the geodemomgraphic characteristics of those involved in pedestrian crashes are most narrow between drivers and pedestrians involved in crashes and living in the lowest deprivation quintiles.

A confounding factor, which might have been specified in advance but is also suggested by the graphics in Fig. 2, is the IMD class of the location in which crashes occur. To explore this, we can return to the observed data (*Phase1-Pass2*) and condition (or facet) on the IMD class of crash location, laying out the faceted plot left-to-right on the ordered IMD classes. Eyeballing this graphic of observed counts (labeled in Fig. 3), we see the heavy association between geodemographics for crashes occurring in the least deprived quintile and elsewhere there is slightly more ‘mixing’. Few pedestrians living outside the most deprived quintile are involved in crashes that occur in that quintile. Given the dominating pattern is of crashes occurring in the most deprived quintiles, however, it is difficult to see too much variation from the diagonal cell in the less-deprived quintiles. An easy adjustment would be to apply a local color scale for each faceted plot – cell counts scaled using the maximum value within the IMD location class – and therefore compare relative “leakage” from the diagonal for each deprivation level of crash location. The more interesting question, however, is whether this known association is stronger for certain driver-pedestrian-location combinations than others: that is, net of the dominant pattern, in which cells are there greater or fewer crash counts, a key objective of *Phase2* (Model + Residual) in our framework.

This is not a straightforward task as the dependency is intrinsic to our contingency table. While previously (*Phase1-Pass1*) we derived a model directly from the contingency table, we now must create a slightly more complex model specification (*Phase2-Pass2*). The concept that we are exploring is whether crash counts vary depending on geodemographic distance of drivers and pedestrians from crash locations. We calculate a new variable measuring this distance (“geodemographic distance”): the euclidean distance between the IMD class of the driver-pedestrian-crash location, treating the IMD class as a continuous variable ranging from 1 to 5. In Fig. 3 we encode this derived variable directly. We then specify a Poisson regression model, modelling crash counts in each driver-pedestrian-crash location cell (y_{ij}) as a function of geodemographic distance ($\beta_1 x_{ij}$) for that cell. Since the range of the crash count varies



Obs: Clear association between IMD class of ped-driver-crash location. Less strong in more deprived locations where greater 'leakage' from diagonals. Generate a geodemographic distance variable — distance between IMD class of ped-driver-crash location.

Exp: Poisson regression modelling counts in each ped-driver-crash location category as a function of geodemographic distance, with varying intercept on IMD class of crash location (due to observed systematic differences in range).

Stat: Signed Pearson residuals. Mathematically equivalent to signed chi-residuals — gives saliency to relative and absolute numbers in obs-exp.

Figure 3. Annotated analysis of pedestrian-driver geodemographics conditioned on crash-location. [Color figure can be viewed at wileyonlinelibrary.com].

systematically by IMD class, identified at *Phase1-Pass1*, we extend the model with a group-level intercept term that varies on the IMD class of the crash location (u_j):

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_j \quad \text{group-level intercept on IMD crash location.} \end{aligned} \tag{3}$$

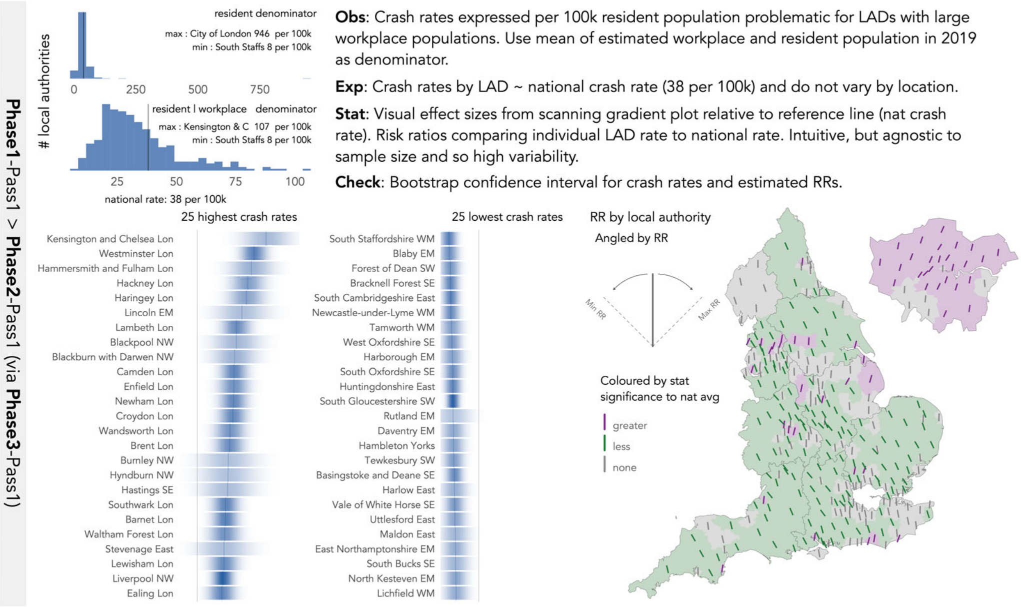
The model diagnostics suggests that geodemographic distance has a strong negative association with cell counts, as expected: higher crash frequencies are recorded in cells of the contingency table that are more geodemographically similar. More interesting is, net of this dominant effect, in which cells of the contingency table the observed data depart most from the model. Residuals from this regression model can be expressed in the same way as in the signed-chi-score model: absolute residuals ($y_{ij} - \exp_{ij}$) are normalized by the square root of the estimated value ($\sqrt{\exp_{ij}}$). The residuals are plotted in the bottom row of Fig. 3. That they are not distributed randomly through the cells of the contingency table suggests that characterizing this difference from expectation may be instructive.

The vertical block of purple in the left column of the left-most matrix (crashes occurring in high-deprivation areas) indicates higher than expected crash counts for pedestrians living and being hit in the most deprived quintile, both by drivers living in that high-deprivation quintile and the less-deprived quintiles (especially the lowest to quintiles), as evidenced by the vertical strip. This pattern is important as it persists even after having modeled for "geodemographic distance." It is replicated to a lesser extent for pedestrians living and being hit in the second most deprived quintile, but does not appear so strongly for the mid- and low-deprivation quintiles. Studying the plots closely there is a tendency for purple cells to appear in the lower half of the cells for matrices describing the mid-low deprivation crash locations. This indicates that, net of geodemographic distance, drivers living in higher deprivation neighborhoods are involved in road crashes with pedestrians in greater number than would be expected given geodemographic distance, and for crashes occurring in the lower deprivation areas in which they do not live.

We should exercise some caution in over-interpreting departures from expectation for the smaller residuals in Fig. 3, although the fact that the residual structure in the tables is nonrandom confers some confidence to this interpretation. Certainly, our analysis reinforces the importance of studying sociodemographic inequalities and road safety, but more uniquely that the combination of sociodemographic characteristics of drivers and pedestrians involved in individual crashes is worthy of special attention. We could preregister some hypotheses to test for the "high deprivation pedestrians in high deprivation areas" effect, although given the fact that we have already used all pedestrian crashes between 2010 and 2019 recorded in STATS19 for which IMD data is available, doing so would risk "double-dipping" (Devezer et al., 2021). More interesting are the numerous lines of inquiry that our model-based analysis provokes. An upshot might be to analyze over more of the detail in the STATS19 data around specific crash characteristics to investigate why it is that both drivers and pedestrians living in higher deprivation areas are overrepresented in the road crash statistics. Also to investigate the "importing effect" indicated by our model at *Phase2-Pass2*: of low-deprivation drivers crashing into high-deprivation pedestrians in high-deprivation areas, and the apparent pattern of higher deprivation drivers crashing into higher deprivation pedestrians even in the less-deprived crash locations (e.g. the residuals plots for IMD class 4 and 5 at *Phase2-Pass2*).

Data Analysis 2: Comparing pedestrian crash rates between areas

A common task in applied road safety analysis is to monitor and compare crash and severity rates between geographic reporting areas (e.g. Jones et al., 2008). Again there is no singular approach to this, with well-documented challenges around the selection of appropriate denominators and in expressing uncertainty due to low counts at the area-level (Eksler, Lassarre, and Thomas, 2008; Eksler, 2010; Boulieri et al., 2017; Scholes et al., 2018). We wish to rank and prioritize geographic areas by LAD, a reporting area frequently used in road safety analysis (Jones et al., 2008). We start by calculating crash rates for pedestrian casualties in 2019 normalized by population size, and express this as the total number of pedestrian casualties by 100,000 population. Population denominators have long been used in road safety research (Smeed, 1949) and are intuitively relevant to pedestrian casualties. Residential population denominators are, though, not always effective measures of "exposure." Distance walked per person per year varies substantially from place to place. Furthermore, for LADs in large urban centers, daytime populations may be orders of magnitude larger than resident populations: crash rates constructed using residential denominators alone may dramatically overstate injury risk in some areas. We confirmed this at *Phase1-Pass1* by generating a histogram over the quantitative crash rate variable and observing that the one-dimensional distribution of crashes is skewed, with extremely high crash rates for LADs in central London. By collecting an estimate of workplace population and using as a denominator the mean of the workplace and residential populations (Fig. 4), we arrive at a


Obs: Crash rates expressed per 100k resident population problematic for LADs with large workplace populations. Use mean of estimated workplace and resident population in 2019 as denominator.

Exp: Crash rates by LAD ~ national crash rate (38 per 100k) and do not vary by location.

Stat: Visual effect sizes from scanning gradient plot relative to reference line (nat crash rate). Risk ratios comparing individual LAD rate to national rate. Intuitive, but agnostic to sample size and so high variability.

Check: Bootstrap confidence interval for crash rates and estimated RRs.

Figure 4. Analysis and ranking of Local Authority Districts by crash rate. [Color figure can be viewed at wileyonlinelibrary.com].

distribution closer to expectation (e.g. Poisson-shaped), with crash rate measured in units of casualties (slight, serious, and fatal) per 100,000 people.

At *Phase2-Pass1*, we make a basic assumption that crash rates for each LAD do not differ from the national crash rate. Before performing any analysis, we acknowledge limitations with this area-level comparison. Although STATS19 is a population dataset to the extent that it contains data on every crash recorded by the police, many LADs will contain relatively few pedestrian injuries; 95 of the 317 LADs in England recorded < 30 pedestrian injuries in 2019, for example. The more data on which our crash rates are based, the more certainty we have in those rates being reliable estimates of crash risk. Before analyzing between-LAD variation in crash rates, we therefore apply statistics and visualization design principles (*Phase3 Infer* + *Check*) so that judgements around LAD-level variation in crash rates at *Phase2-Pass1* are made while accounting visually for variable levels of uncertainty. Our observed data is treated as a sample of an (unobtainable) population and crash rates are parameters that try to represent, or estimate, this population. From here we can apply some statistical concepts to quantify uncertainty around the calculated crash rates: that a sampling distribution can be derived and used to quantify the precision of each crash rate estimate. This sampling distribution is generated empirically via a non-parametric bootstrap consisting of 1,000 resamples with replacement. Upper and lower limits are lifted from 0.025 and 0.975 percentile positions of the bootstrap sampling distribution. In Fig. 4, gradient bars (Correll and Gleicher, 2014) are used to show the range in values from this sampling distribution for each LAD, focusing on the top and bottom LADs in terms of crash rate with thin vertical lines positioned at the center of this distribution (the estimated crash rate). This encoding invites us to think distributionally and consider the uncertainty of our estimated rates (Padilla, Kay, and Hullman, 2021). Note that the uncertainty ranges, and their visual representation as gradient plots, is a *Phase3* class of activity. We are using *infer* + *check* type techniques to discourage false discovery early in the data analysis (*Phase2-Pass1*). Also in Fig. 4 is a map of these crash rates expressed as risk ratios, where the crash rate for each LAD is compared with the crash rate expected by the England average (38 per 100,000 people).

The gradient plots in Fig. 4 suggest high uncertainty (range in bootstrap) especially for LADs identified as having the highest crash rates. This is due to crash rates being derived from a comparatively small number of observations. If this ranking is to be used to prioritize research and policy attention, then we may wish to apply techniques to address this – for example, biasing crash rates toward the global mean (national average rate) where they are based on smaller numbers of observations. A clear finding is that London boroughs dominate amongst the highest crash rates. This suggests our population-based denominator, which tries to estimate exposure, may be insufficient for dense urban LADs which "import" people other than residents and workers in large numbers. Alternative denominators for estimating the population at risk may be sought, or alternatively we may wish to adjust risk estimates taking into account population density, used elsewhere as an umbrella variable for capturing distinct urban-rural contexts (e.g. Eksler, Lassarre, and Thomas, 2008).

The map of risk ratios suggests heavy spatial autocorrelation in crash rates. Each LAD is represented as a line-angled according to how much greater (angled to the right) or lesser (angled to the left) the crash rate in that area is relative to the national average. Lines are further colored by whether those crash rates are statistically different from the national average – whether the bootstrap interval of the RRs crosses 1.0. Orientation is applied widely in cartography and there are useful gestalt properties that help characterize the dominant pattern of spatial autocorrelation in crash rates (similarly oriented lines), but also isolate LADs which are locally exceptional.

In *Phase2-Pass2* we update the model used to estimate crash rates. This is to address the problems identified at *Phase2-Pass1*: of generating reliable crash estimates for LADs with small numbers of observations and accounting for some of the unique context in dense urban LADs. To do this we specify a hierarchical model where the adjusted crash rates by LAD ($y_i ... y_n$) are assumed to be Poisson-distributed and conditioned on population density (represented as a fixed effect β_{1i}) and a random effect offset on LAD v_i. The population density term is assumed to represent an umbrella for local context – local differences in mobility needs and road-user mix, demographic composition, level of urbanization, the structure and quality of the infrastructure and emergency services response times (Eksler, Lassarre, and Thomas, 2008; Eksler, 2010). The random effect on LAD determines how much crash rates with small counts are shrunk to the global mean. The exponential of this random effect ($exp\left(u_i\right)$) is a RR, also called Bayes relative risk, similar to those presented in Fig. 4, but adjusted for population density and LAD-level sample size:

$$y_i = \beta_{0i} + \beta_{1i} + \varepsilon_i$$

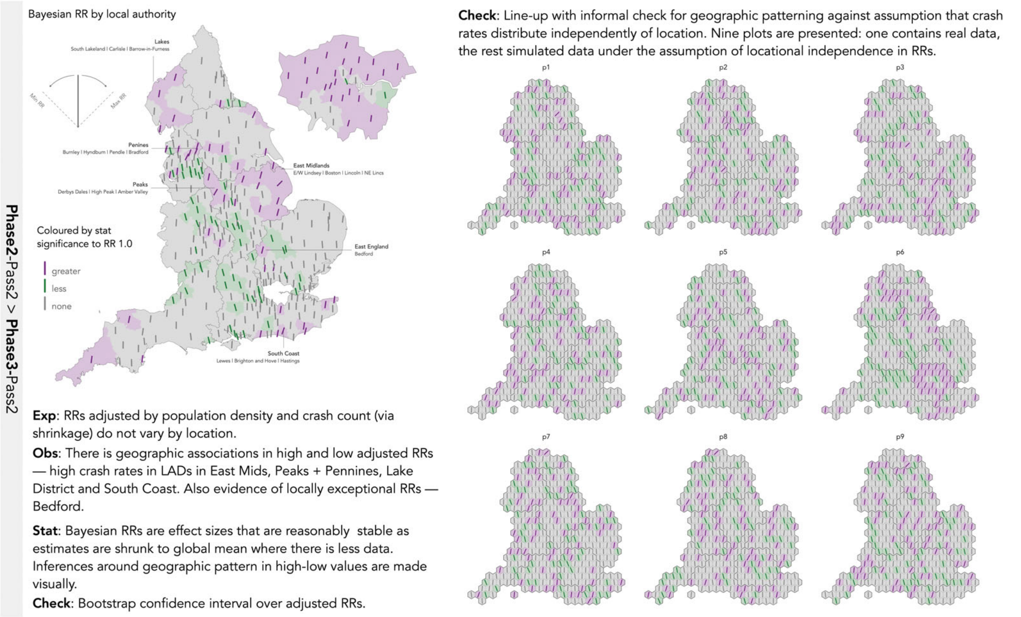
$$\beta_{0i} = \beta_0 + u_i \quad \text{random intercept LAD}$$

y_i estimated crash rate for an individual LAD, conditioned on:

β_{0i} a random effect term for each LAD (i);

ε_{1i} a fixed effect recording the log population density of each LAD (i). (4)

Our new rankings, based on an updated and slightly more complex model formulation, support comparison and area-level prioritization that is more likely to be robust to replication


Phase2-Pass2 > Phase3-Pass2
Bayesian RR by local authority
Min RR
Max RR
Lakes
South Lakeland | Carlisle | Barrow-in-Furness
Penines
Burnley | Hyndburn | Pendle | Bradford
Peaks
Derbys Dales | High Peak | Amber Valley
East Midlands
E/W Lindsey | Boston | Lincoln | NE Lincs
East England
Bedford
South Coast
Lewes | Brighton and Hove | Hastings
Coloured by stat significance to RR 1.0
greater
less
none
p1
p2
p3
p4
p5
p6
p7
p8
p9

Exp: RRs adjusted by population density and crash count (via shrinkage) do not vary by location.

Obs: There is geographic associations in high and low adjusted RRs — high crash rates in LADs in East Mids, Peaks + Pennines, Lake District and South Coast. Also evidence of locally exceptional RRs — Bedford.

Stat: Bayesian RRs are effect sizes that are reasonably stable as estimates are shrunk to global mean where there is less data. Inferences around geographic pattern in high-low values are made visually.

Check: Bootstrap confidence interval over adjusted RRs.

Check: Line-up with informal check for geographic patterning against assumption that crash rates distribute independently of location. Nine plots are presented: one contains real data, the rest simulated data under the assumption of locational independence in RRs.

Figure 5. Analysis of Local Authority District (LAD) crash rates with RRs adjusted by population density and sample size. Hex cartogram layout defined by ODI Leeds's `HexJson` format (ODI Leeds, 2021). [Color figure can be viewed at wileyonlinelibrary.com].

(due to the shrinkage term) and which via the population density variable captures some of the contextual bias suggested by the *Phase2-Pass1* analysis.

In Fig. 5 these updated RRs are mapped in the same way as in Fig. 4. Although it remains exceptional, there are now slightly fewer local authorities in London identified as having RRs statistically > 1.0 and the effect sizes (line angles) are also slightly dampened. There are fewer authorities identified as having RRs statistically < 1.0; 143 in the original analysis versus 60 in the updated model. The remaining groups of LADs with RRs statistically different from 1.0 appear to be geographically concentrated. Those with elevated RRs are annotated in Fig. 5: sets of authorities on the South Coast, East Midlands, Peaks and Pennines and North West/Lake District.

Previously we observed that our unadjusted RRs are heavily spatially autocorrelated. This is to be expected in most geographic phenomena (Tobler, 1970). In conditioning on population density, however, the autocorrelation effect in the adjusted RRs in Fig. 5 is very much dampened. As the model includes greater local context, we might hope that the spatial pattern in RRs is closer to random. We could test for this by calculating global autocorrelation statistics and comparing across the two RR datasets. Instead we generate a graphical line-up test (Buja et al., 2009) where the observed data is hidden among a set of decoy plots of the same data, but with RR values randomly permuted across local authorities. Such tests have particular utility in geographical analysis (Widen et al., 2016; Beecham et al., 2021), as there is a tendency to over-interpret geographic structure represented in map designs (Klippel, Hardisty, and Li, 2011; Doppler et al., 2021). Since the docoy plots are constructed based on a random allocation, the line-up test in Fig. 5 might be regarded as a visual equivalent of a null hypothesis test against locational independence in RRs. We use it, however, to perform a slightly different type of reasoning which closer approximates to the priorities of decision-makers wishing to identify geographic areas for policy attention: whether it is possible that the sorts of groupings annotated in the observed data are genuine ones, or whether they are a result of chance process and induced by the way in which the graphic is composed.

Since the standard geographic representation of England is highly familiar, and we have already seen the real data, the line-up is constructed using an abstracted geographic layout. Each LAD is represented as a hexagon of constant size and with an approximate spatial arrangement that tries to preserve adjacency relations. As well as reducing recognizability, map layouts with spatial units of regular size have been demonstrated empirically to be more effective than real geographic layouts at supporting inferences around spatial dependency (Beecham et al., 2017). From our visual inspection of the line-up in Fig. 5 we can correctly identify the real dataset from the decoys (*p6*) based on the fact that change in line angle is smoother and more structured in this plot and there is more grouping in colors (RRs that are statistically > or < 1.0). Partly this is due to the visual saliency of the statistically elevated RRs in London. However, additional scanning across the real plot and the reference or decoy plots enables informal checks (*P3 Infer + Check*) on groupings outside of London. In the decoys it is possible to detect apparent grouping of high and low values that happen by chance, but the high RR groupings for local authorities in the North West, Pennines and East Midlands do appear particularly salient when compared against these reference (decoy) plots. So this *Phase3-Phase2* helps informally check and lend greater confidence to inferences around priority areas that might not be so easily specified and communicated through standard model formulations.

Discussion

The model-based analysis described above may be typical of many geographical analyses of observational data. However, in designing a framework for visually supported exploratory analysis along the lines articulated by Hullman and Gelman (2021b) and others (Heer, 2021), its presentation and positioning is distinctive and worthy of discussion. In this section we reflect on the two data analyses to make three claims about the usefulness of the framework for advancing and injecting rigor into modern geographical analysis.

Our framework progresses a data analysis by supporting model refinement

Exploratory analysis is typically presented as a winnowing-type activity. Statistical and graphical summaries are generated from various data processing operations and the insights used to delimit the scope for follow-up analysis. This narrowing of activity in our framework applies to the process of model-building. At the very early stages of analysis, models may be naive and embedded with limited local context (information). As the analysis progresses and new knowledge is developed, models are updated with more information. Our data graphics provide an interface to this activity and, as demonstrated in other applied visual data analyses (Wood et al., 2011; Beecham and Slingsby, 2019; Beecham, Williams, and Comber, 2020), the configurations of the graphics themselves help to suggest model updates.

As an example, Data Analysis 1 (summarized in Fig. 6) was heavily data-driven. The contingency tables on which the analysis was built were generated from abstracting over relevant data at *Phase1-Pass1* – identifying the three categorical-ordinal IMD variables and laying these out in ordered heatmap matrices. When interpreting these matrices, we regarded them as categorical-ordinal equivalents of scatterplots and so consciously “looked for” linear

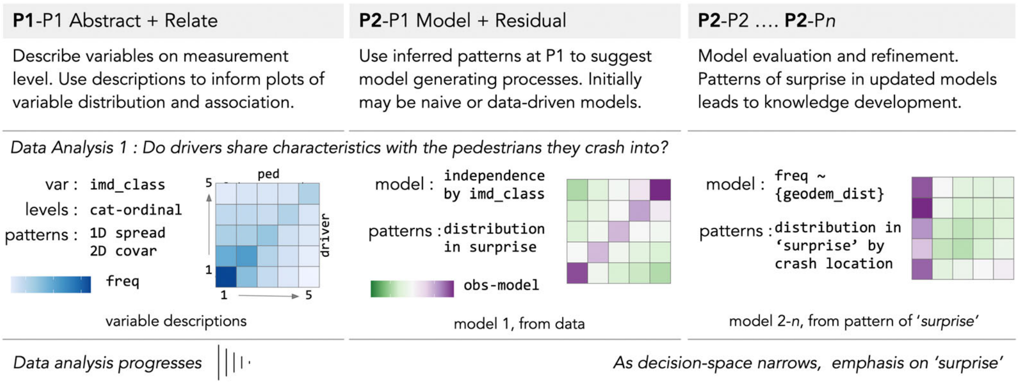

P1-P1 Abstract + Relate
Describe variables on measurement level. Use descriptions to inform plots of variable distribution and association.
P2-P1 Model + Residual
Use inferred patterns at P1 to suggest model generating processes. Initially may be naive or data-driven models.
P2-P2 .... P2-Pn
Model evaluation and refinement. Patterns of surprise in updated models leads to knowledge development.
Data Analysis 1 : Do drivers share characteristics with the pedestrians they crash into?
var : imd_class
levels : cat-ordinal
patterns : 1D spread
2D covar
freq
ped
driver
5
1
1
5
variable descriptions
model : independence by imd_class
patterns : distribution in surprise
obs-model
model 1, from data
model : freq ~ {geodem_dist}
patterns : distribution in 'surprise' by crash location
model 2-n, from pattern of 'surprise'
Data analysis progresses
As decision-space narrows, emphasis on 'surprise'

Figure 6. Data Analysis 1 located within our framework. [Color figure can be viewed at wileyonlinelibrary.com].

association. The subsequent model of no association could also be understood as data-driven since expected values were derived directly from the contingency table layout. Encoding these expected values within the heatmap matrices at *Phase2-Pass1* (Fig. 2) and comparing across the observed, expected and residual plots, lent intuition to how expectation is spread over the contingency table under the assumption of independence in pedestrian-driver geodemographics. The strong diagonals in these plots then provided justification for conditioning on crash location at *Phase1-Pass2* and from here the proposal for "geodemographic distance" as an explanatory variable in the full regression model at *Phase2-Pass2*.

Our framework emphasizes "surprise" – characterizing difference from expectation in detail

A feature of both analyses was that, rather than building the analysis around standard model diagnostics – global chi-square statistics, regression coefficients and model fits – most effort was concentrated on the patterning of model residuals. In Data Analysis 1 (Section 5.1 and Fig. 6) contingency tables were used to explore variation in the geodemographic characteristics of pedestrians and vehicles involved in road crashes. That there was an association between the IMD class of pedestrians, drivers and crash locations was less interesting than *where* in the contingency table this effect was particularly concentrated, as this suggests specific combinations of driver-pedestrian-location characteristic for further investigation. In Data Analysis 2 (Section 5.2 and Fig. 7), where the objective was to prioritize LADs and identify geographic groupings with exceptional crash rates, a natural focus was on model residuals, the RRs in this case. Comparison to expectation was intrinsic to the choice of encodings in our graphics: lines oriented relative to expectation, colored by direction of statistical effect and with a spatial arrangement to emphasize geographical continuities and discontinuities in RR values.

Close attention to model residuals is commonplace in geographical analysis as it informs the specification of updates that capture spatial dependency and heterogeneity in process. An extension to Data Analysis 2 might be to update our Poisson-regression model with such explicit space effects, as do Boulieri et al. (2017). Graphical approaches used to characterize and check difference from expectation could equally apply to any updated specification. The purpose, according to our framework, would be to generate and express more detailed knowledge relevant to the analysis need, in this case identifying with confidence geographic areas for policy attention.

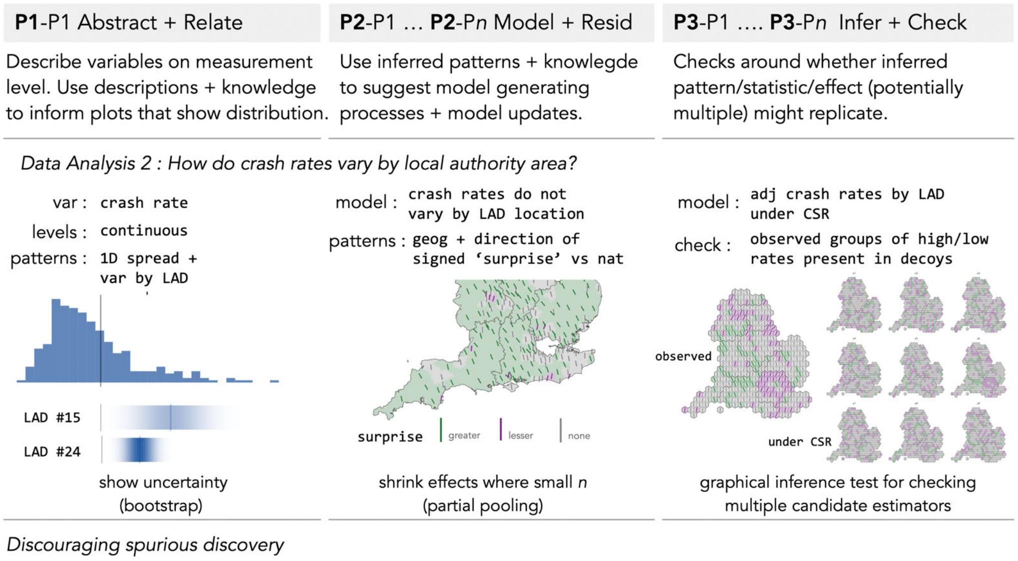

P1-P1 Abstract + Relate
Describe variables on measurement level. Use descriptions + knowledge to inform plots that show distribution.
P2-P1 … P2-Pn Model + Resid
Use inferred patterns + knowlegde to suggest model generating processes + model updates.
P3-P1 …. P3-Pn Infer + Check
Checks around whether inferred pattern/statistic/effect (potentially multiple) might replicate.
Data Analysis 2 : How do crash rates vary by local authority area?
var : crash rate
levels : continuous
patterns : 1D spread + var by LAD
LAD #15
LAD #24
show uncertainty (bootstrap)
model : crash rates do not vary by LAD location
patterns : geog + direction of signed 'surprise' vs nat
surprise
greater
lesser
none
shrink effects where small n (partial pooling)
model : adj crash rates by LAD under CSR
check : observed groups of high/low rates present in decoys
observed
under CSR
graphical inference test for checking multiple candidate estimators
Discouraging spurious discovery

Figure 7. Data Analysis 2 located within our framework. [Color figure can be viewed at wileyonlinelibrary.com].

Such an emphasis may again be familiar to geographical analysis, but it is instructive to position the characterization of "surprise" as an explicit model-building goal, not solely the pursuit of an optimal model formulation judged on parsimony and fit diagnostics.

Our framework discourages spurious discovery

A worthy charge against exploratory visual analysis approaches is that they are vulnerable to over-interpretation and false discovery. Data graphics emphasize data patterns, but the complexities around those patterns may be overlooked (Hullman and Gelman, 2021b). This applies especially to road safety research as it is often necessary to condition on detailed contextual variables and therefore work with data subsets of diminishing sample size. Our framework makes explicit the need to *infer + check* observed patterns in data and models. In Data Analyses 1 and 2 this is achieved first through our use of quantitative measures. The chi-square residuals and Pearson-regression residuals give greater salience to differences from expectation that are large in both relative and absolute number. In the regression model developed in Data Analysis 2, we use partial pooling to bias LAD-level crash rates to the global mean (national average rate) depending on the number of observations on which they are based. Where possible we also embed skepticism into our visualization design. The gradient bars in Data Analysis 2 (Fig. 4), for example, prevent against over-interpretation of area-level rankings in estimated crash rates, especially for LADs positioned amongst the highest crash rates in the country. There is a growing repertoire of empirically tested techniques for uncertainty representation enumerated in our framework (Fig. 1), and demonstrated in this paper's accompanying code repository, and for which there are software libraries to support easy implementation (cf. Kay, 2021a; Kay, 2021b).

In attempting to formalize exploratory visual data analysis Hullman and Gelman (2021b) invoke the idea of graphical inference as model check (Gelman, 2004), where observed data are compared graphically to reference data replicated under a model. Rather than constraining graphical inference to narrow null hypothesis tests, Hullman and Gelman (2021a) position graphical model checks within a Bayesian framework. This proposal accepts flexibility around the features in the observed data being tested and that expectation will be contingent on analysts' prior knowledge and experience. We find this compelling for geographical analysis as immediate statistical tests at the disposal of spatial analysts, for example, global and local indicators of spatial autocorrelation, do not always map directly to the sorts of observations being made and therefore requiring scrutiny. Data Analysis 2 demonstrates this. In Fig. 5 we used a graphical line-up test (Buja et al., 2009) with decoy plots that assume complete spatial randomness. Rather than treating this as a standard graphical inference test where inferences are limited to statements relative to a null hypothesis of complete spatial randomness, we treated specific features in the observed plot itself as statistical estimators/identifiers (Cook, Reid, and Tanaka, 2021) and compared how those estimators behave in the repeated samples (the decoys). This approach does not allow formal statements about knowledge in light of evidence to be made, but does lend some rigor to the sorts of detailed (multimodal) spatial patterns in the observed data that are of direct interest to road safety analysts and practitioners.

Conclusion

The starting point for this paper was around the difficulties in working with a specific dataset: STATS19 road crash data. It is possible to derive many interesting patterns from detailed information on crash context and the vehicles and individuals involved. These are nevertheless patterns in which we have varying confidence. Uncertainty around datasets, the selection of context variables and of appropriate statistical techniques and reporting mechanisms, means it is difficult to claim empirical knowledge using the sorts of research designs required by the Scientific Reform movement (Devezer et al., 2021): *EDA → preregistration → CDA*. This difficulty is not confined to road safety analysis. In modern geographic research, messy observational data are repurposed to investigate spatial processes that are often fluidly defined (Miller and Goodchild, 2015; Singleton and Arribas-Bel, 2021). Although there are calls for more formal and incremental knowledge development in quantitative geography, issues of inference, generalization and replicability of findings are still to be resolved (Kedron et al., 2021; Wolf et al., 2021). Separately Hullman and Gelman (2021b), taking a critical view of the EDA-CDA dichotomy, initiate a lively discussion in how exploratory visual analysis approaches might be formalized. We present a framework, and implemented analysis, that attempts to reify this rethinking through data analysis phases which emphasize model-building and evaluation, and which presents techniques for addressing issues of inference and false discovery. That the framework is instantiated through a detailed data analysis is an important contribution. We demonstrate practically how greater rigor can be injected into the sorts of exploratory visual analysis activities that remain widespread, and necessary, to modern geographical analysis.

References

Aldred, R., A. Goodman, J. Gulliver, and J. Woodcock. (2018). "Cycling Injury Risk in London: A Case-Control Study Exploring the Impact of Cycle Volumes, Motor Vehicle Volumes, and Road Characteristics Including Speed Limits." *Accident Analysis & Prevention* 117, 75–84.

Amrhein, V., D. Trafimow, and S. Greenland. (2019). "Inferential Statistics as descriptive statistics: There Is No Replication Crisis if We Don't Expect Replication." *The American Statistician* 73, 262–70.

Beck, L. F., A. M. Dellinger, and M. E. O'Neil. (2007). "Motor Vehicle Crash Injury Rates by Mode of Travel, United States: Using Exposure-Based Methods to Quantify Differences." *American Journal of Epidemiology* 166, 212–8.

Beecham, R. (2022). Supplementary Materials for the Paper "A Framework for Inserting Visually-Supported Inferences Into Geographical Analysis Workflow: Application to Road Crash Analysis". https://github.com/rogerbeecham/vis-inferences-roadsafety.

Beecham, R., J. Dykes, L. Hama, and N. Lomax. (2021). "On the Use of 'Glyphmaps' for Analysing COVID-19 Reported Cases." *ISPRS International Journal of Geo-Information* 10, 213.

Beecham, R., J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood. (2017). "Map LineUps: Effects of Spatial Structure on Graphical Inference." *IEEE Transactions on Visualization and Computer Graphics* 23, 391–400.

Beecham, R., and A. Slingsby. (2019). "Characterising Labour Market Self-Containment in London with Geographically Arranged Small Multiples." *Environment and Planning A: Economy and Space* 51, 1217–24.

Beecham, R., N. Williams, and L. Comber. (2020). "Regionally-Structured Explanations Behind Area-Level Populism: An Update to Recent Ecological Analyses." *PLoS One* 15, e0229974.

Boss, D., T. Nelson, and M. Winters. (2018). "Monitoring City Wide Patterns of Cycling Safety." *Accident Analysis & Prevention* 111, 101–8.

Boulieri, A., S. Liverani, K. de Hoogh, and M. Blangiardo. (2017). "A Space-Time Multivariate Bayesian Model to Analyse Road Traffic Accidents by Severity." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 119–39.

Brunsdon, C. (2014). "Spatial Science – Looking Outward." *Dialogues in Human Geography* 4, 45–9.

Brunsdon, C. (2016). "Quantitative Methods I: Reproducible Research and Quantitative Geography." *Progress in Human Geography* 40, 687–96.

Buja, A., D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. (2009). "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Royal Society Philosophical Transactions A* 367, 4361–83.

Cook, D., N. Reid, and E. Tanaka. (2021). "The Foundation Is Available for Thinking About Data Visualization Inferentially." *Harvard Data Science Review 3(3).* https://hdsr.mitpress.mit.edu/pub/mpdasaqt

Correll, M., and M. Gleicher. (2014). "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error." *IEEE Transactions on Visualization and Computer Graphics* 20, 2142–51.

Correll, M., and J. Heer. (2017). "Surprise! Bayesian Weighting for De-Biasing Thematic Maps." *IEEE Transactions on Visualization & Computer Graphics* 23, 651–60.

DefRA. (2018) Rural deprivation. https://www.gov.uk/government/statistics/rural-deprivation-statistics.

Devezer, B., D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas. (2021). "The Case for Formal Methodology in Scientific Reform." *Royal Society Open Science* 8, 200805.

Doppler, J. H., M. Pohl, R. Beecham, and J. Dykes. (2021). "Strategies for Detecting Difference in Map Line-Up Tasks." In *INTERACT 2021 - 18th International Conference on Human-Computer Interaction, Bari, Italy, August 30th-2nd September, 2021, Proceedings*. *Lecture Notes in Computer Science*, Vol 12934, edited by D. Lamas, F. Loizides, L. E. Nacke, H. Petrie, M. Winckler, and P. Zaphiris. Cham, Switzerland: Springer.

Eksler, V. (2010). "Measuring and Understanding Road Safety Performance at Local Territorial Level." *Safety Science* 48, 1197–202.

Eksler, V., S. Lassarre, and I. Thomas. (2008). "Regional Analysis of Road Mortality in Europe." *Public Health* 122, 826–37.

Elragal, A., and R. Klischewski. (2017). "Theory-Driven or Process-Driven Prediction? Epistemological Challenges of Big Data Analytics." *Journal of Big Data* 4, 19.

Elvik, R., T. Vaa, A. Erke, and M. Sorensen. (2009). *The Handbook of Road Safety Measures*. Bingley, UK: Emerald Group Publishing.

Feleke, R., S. Scholes, M. Wardlaw, and J. S. Mindell. (2018). "Comparative Fatality Risk for Different Travel Modes By Age, Sex, and Deprivation." *Journal of Transport & Health* 8, 307–20.

Gelman, A. (2004). "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13, 755–79.

Goodchild, M. F., and W. Li. (2021). "Replication Across Space and Time Must Be Weak In the Social and Environmental Sciences." *Proceedings of the National Academy of Sciences* 118, e2015759118.

Graser, A. (2021). "An Exploratory Data Analysis Protocol for Identifying Problems in Continuous Movement Data." *Journal of Location Based Services* 15, 89–117.

Heer, J. (2021). "Exploratory Analysis and Its Malcontents." *Harvard Data Science Review, 3(3)*. https://hdsr.mitpress.mit.edu/pub/vszs87oj

Hindle, G. A., T. Hindle, and S. Souli. (2009). "Modelling and Assessing Local Area Differences In Road Casualties: A Case Study in England." *Journal of the Operational Research Society* 60, 781–8.

Hullman, J., and A. Gelman. (2021a). "Challenges in Incorporating Exploratory Data Analysis Into Statistical Workflow." *Harvard Data Science Review*, 3(3). https://hdsr.mitpress.mit.edu/pub/2ym7zm34

Hullman, J., and A. Gelman. (2021b). "Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference." *Harvard Data Science Review*, 3(3). https://hdsr.mitpress.mit.edu/pub/w075glo6

Hullman, J., P. Resnick, and E. Adar. (2015). "Hypothetical Outcome Plots Outperform Error Bars And Violin Plots for Inferences About Reliability of Variable Ordering." *PLoS One* 10, e0142444.

Jones, A., R. Haynes, V. Kennedy, I. Harvey, T. Jewell, and D. Lea. (2008). "Geographical Variations in Mortality and Morbidity From Road Traffic Accidents in England and Wales." *Health & Place* 14, 519–35.

Kale, A., F. Nguyen, M. Kay, and J. Hullman. (2019). "Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data." *IEEE Transactions on Visualization and Computer Graphics* 25, 892–902.

Kay, M. (2021a). *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.0.1. https://mjskay.github.io/ggdist/

Kay, M. (2021b). *Tidybayes: Tidy Data and Geoms for Bayesian Models*. R package version 3.0.0. http://mjskay.github.io/tidybayes/.

Kedron, P., A. E. Frazier, A. B. Trgovac, T. Nelson, and A. S. Fotheringham. (2021). "Reproducibility and Replicability in Geographical Analysis." *Geographical Analysis* 53, 135–47.

Kitchin, R. (2014). "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1, 1–12.

Klippel, A., F. Hardisty, and R. Li. (2011). "Interpreting Spatial Patterns: an Inquiry Into Formal and Cognitive Aspects of Tobler's First Law of Geography." *Annals of the Association of American Geographers* 101, 1011–31.

Kondo, M. C., C. Morrison, E. Guerra, E. J. Kaufman, and D. J. Wiebe. (2018). "Where Do Bike Lanes Work Best? A Bayesian Spatial Model of Bicycle Lanes and Bicycle Crashes." *Safety Science* 103, 225–33.

Liu, L., A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. (2017). "Uncertainty Visualization By Representative Sampling From Prediction Ensembles." *IEEE Transactions on Visualization and Computer Graphics* 23, 2165–78.

Loidl, M., C. Traun, and G. Wallentin. (2016) Spatial Patterns and Temporal Dynamics of Urban Bicycle Crashes: A Case Study From Salzburg (Austria). *Journal of Transport Geography*, 52, 38–50. https://www.sciencedirect.com/science/article/pii/S0966692316000302.

Lovelace, R., M. Morgan, L. Hama, M. Padgham, D. Ranzolin, and A. Sparks. (2019). "stats 19: A Package for Working with Open Road Crash Data." *The Journal of Open Source Software* 4, 1181.

Lovelace, R., H. Roberts, and I. Kellar. (2016). "Who, Where, When: The Demographic and Geographic Distribution of Bicycle Crashes in West Yorkshire." *Transportation Research Part F: Traffic Psychology and Behaviour* 41, Part B, 277–293.

McIntosh, R. D. (2017). "Exploratory Reports: A New Article Type for Cortex." *Cortex* 96, A1–4.

Miller, H. J., and M. F. Goodchild. (2015). "Data-Driven Geography." *GeoJournal* 80, 449–61.

Morris, T. P., I. R. White, and M. J. Crowther. (2019). "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38, 2074–102.

Noble, S., McLennan, D., Noble, M., Plunkett, E., Gutacker, N., Silk, M. and Wright, G. (2019) The English Indices of Deprivation 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.

ODI Leeds. (2021). HexJSON Format. https://open-innovations.org/projects/hexmaps/hexjson. [Online; accessed 20-December-2021].

Open Science Collaboration. (2015). "Estimating the Reproducibility of Psychological Science." *Science* 349, aac4716.

O'Toole, S. E., and N. Christie. (2018). "Deprivation and Road Traffic Injury Comparisons for 4–10 and 11–15 Year-Olds." *Journal of Transport & Health* 11, 221–9.

Padilla, L., M. Kay, and J. Hullman. (2021). "Uncertainty Visualization." In *Wiley StatsRef: Statistics Reference Online*, edited by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. New Jersey: Wiley.

Quinan, P. S., and M. Meyer. (2016). "Visually Comparing Weather Features in forecasts." *IEEE Transactions on Visualization and Computer Graphics* 22, 389–98.

Sanders, R. L. (2015). "Perceived Traffic Risk For Cyclists: The Impact of Near Miss and Collision Experiences." *Accident Analysis & Prevention* 75, 26–34.

Sarkar, C., C. Webster, and S. Kumari. (2018). "Street Morphology and Severity of Road Casualties: A 5-Year Study of Greater London." *International Journal of Sustainable Transportation* 12, 510–25.

Scholes, S., M. Wardlaw, P. Anciaes, B. Heydecker, and J. S. Mindell. (2018). "Fatality Rates Associated with Driving and Cycling for All Road Users in Great Britain 2005–2013." *Journal of Transport & Health* 8, 321–33.

Singleton, A., and D. Arribas-Bel. (2021). "Geographic Data Science." *Geographical Analysis* 53, 61–75.

Smeed, R. J. (1949). "Some Statistical Aspects of Road Safety Research." *Journal of the Royal Statistical Society. Series A (General)* 112, 1–34.

Stevens, S. (1946). "On the Theory of Scales of Measurement." *Science* 103, 677–80.

Szollosi, A., and C. Donkin. (2021). "Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research." *Perspectives on Psychological Science* 16, 717–24.

Tobler, W. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46, 234–40.

Tortosa, E. V., R. Lovelace, E. Heinen, and R. P. Mann. (2021). "Socioeconomic Inequalities in Cycling Safety: An Analysis of Cycling Injury Risk by Residential Deprivation Level in England." *Journal of Transport & Health* 23, 101291.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Visalingam, M. (1981). "The Signed Chi-Score Measure for the Classification and Mapping of Plychotomous Data." *The Cartographic Journal* 18, 32–43.

Wickham, H., D. Cook, H. Hofmann, and A. Buja. (2010). "Graphical Inference for Infovis." *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2019)* 16, 973–9.

Widen, H. M., J. B. Elsner, S. Pau, and C. K. Uejio. (2016). "Graphical Inference in Geographical Research." *Geographical Analysis* 48, 115–31.

Winters, M., G. Davidson, D. Kao, and K. Teschke. (2011). "Motivators and Deterrents of Bicycling: Comparing Influences on Decisions to Ride." *Transportation* 38, 153–68.

Wolf, L. J., S. Fox, R. Harris, R. Johnston, K. Jones, D. Manley, E. Tranos, and W. W. Wang. (2021). "Quantitative Geography III: Future Challenges and Challenging Futures." *Progress in Human Geography* 45, 596–608.

Wood, J., D. Badawood, J. Dykes, and A. Slingsby. (2011). "BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers." *IEEE Transactions on Visualization and Computer Graphics* 17, 2384–91.