Predictive spatial network analysis for high-resolution transport modeling, applied to cyclist flows, mode choice, and targeting investment

Crispin H. V. Cooper

Published online: 12 Mar 2018.

Submit your article to this journal ⬀

Article views: 1620

View related articles ⬀

View Crossmark data ⬀

Taylor & Francis
Taylor & Francis Group

∂ OPEN ACCESS   Check for updates

Predictive spatial network analysis for high-resolution transport modeling, applied to cyclist flows, mode choice, and targeting investment

Crispin H. V. Cooper

Sustainable Places Research Institute, Cardiff University, Cardiff, Wales, United Kingdom

ABSTRACT

Betweenness is a measure long used in spatial network analysis (SpNA) to predict flows of pedestrians and vehicles, and more recently in public health research. We improve on this approach with a methodology for combining multiple betweenness computations using cross-validated ridge regression to create wide-scale, high-resolution transport models. This enables computationally efficient calibration of distance decay, agglomeration effects, and multiple trip purposes. Together with minimization of the Geoffrey E. Havers (GEH) statistic commonly used to evaluate transport models, this bridges a gap between SpNA and mainstream transport modeling practice. The methodology is demonstrated using models of bicycle transport, where the higher resolution of the SpNA models compared to mainstream (four-step) models is of particular use. Additional models are developed incorporating heterogeneous user preferences (cyclist aversion to motor traffic). Based on network shape and flow data alone the best model gives reasonable correlation against cyclist flows on individual links, weighted to optimize GEH ($r^2 = 0.78$, GEH = 1.9). As SpNA models use a single step rather than four, and can be based on flow data alone rather than demographics and surveys, the cost of calibration is lower, ensuring suitability for small-scale infrastructure projects as well as large-scale studies.

1. Introduction

1.1 Comparison between four-step models and spatial network analysis

Betweenness is a measure of spatial network accessibility invented by Freeman (1977) and is used in modern social network analysis (SNA) and spatial network analysis (SpNA). It has a history of use in the production of models to fit pedestrian and vehicle flows (Cooper, 2015; Haworth, 2014; Hillier & Iida, 2005; Jayasinghe, 2017; Lowry, 2014; Omer et al., 2017; Patterson, 2016; Serra & Hillier, 2017; Turner, 2007) but is not used in mainstream motor vehicle transport modeling for which the four-step model (Ortúzar & Willumsen, 2011) is ubiquitous. Due to their simplified nature, SpNA models have also been used in epidemiology to quantify built environment factors for individuals (Cooper, Fone, & Chiaradia, 2014; Fone et al., 2012; Sarkar et al., 2015; Sarkar, Gallacher, & Webster, 2013; Sarkar, Webster, & Gallacher, 2014).

The aim of this article is not to challenge the highly developed four-step model, but to present a nascent alternative methodology based on the combination of SpNA with cross-validated regression techniques which can handle collinear predictors. To demonstrate the methodology we focus on bicycle transport, for which models are currently lacking, and for

which four-step models are not currently used as they are typically not of sufficient resolution to capture cyclist behavior.

Four-step models, as the name suggests, model transport in four stages: trip generation, distribution, mode choice, and assignment. Generation models are typically based on demographic data, and distribution models predict trips from homes to destinations such as work places, retail facilities, and other homes. The mode choice model predicts the split of travel choices (private car, public transport, etc.) and finally the assignment model predicts the actual route taken, given an origin, destination, and mode of transport. This is a simplified picture as various feedbacks between the stages exist (e.g., the least cost route in the assignment model affects the first three stages).

The usual SpNA approach, by contrast, is to skip the first three phases and jump straight to an assignment model (betweenness; described further in Section 2.2). This model uses "everywhere" as both an origin and destination subject to appropriate distance constraints for the mode of travel under consideration. Thus there is a contrast between four-step models, in which links in the network serve only to carry traffic from origin to destination zones, and SpNA models, where there are no zones and the links themselves also fulfil the roles of origin and destination.

To a transport practitioner, it should seem imprudent to model indiscriminate trips from everywhere to everywhere

---

when this is clearly not what happens in reality. The SpNA practitioner would defend their approach as follows. (1) Doing so gives a reasonable level of correlation with flows. In the short term, four-step models give better predictions but in the long term they lose accuracy as land use change occurs. (2) The premise of SpNA is to model the effect of the network itself on flows, as the network is the slowest aspect of the urban environment to change, with land use intensifying in more accessible areas (Chiaradia, Cooper, & Wedderburn, 2014). If network link density is taken as a proxy for land use intensity, standard betweenness implicitly incorporates an elastic model of demand with respect to opportunity: build more network, and more trips will occur, because of both land use intensification and individual response to increased supply (Cooper, 2017) as well as the increase in accessibility of existing destinations considered by four-step models.

Admittedly however there are some problems with this defense. Firstly, it ignores historical land use; although in some cases there is a strong linear relationship between land use and network density (Chiaradia et al., 2014) this is not always the case. Second, standard betweenness assumes fully elastic demand for all areas, while four-step models do not. In reality some areas are more elastic than others: due to agglomeration effects, increasing the accessibility of a city center will generate more additional travel demand than increasing the accessibility of anywhere else. Thirdly, standard betweenness tends to ignore distance decay, preferring instead a sharp cut-off for the maximum trip distance; again this is usually unrealistic (Gao, Wang, Gao, & Liu, 2013) though in occasional circumstances appropriate for cycling (Wardman, Tight, & Page, 2007).

These points illustrate a technical gap between four-step models and SpNA. There is also a practical gap: the former requires high modeling effort (using four models rather than one), strives for high short-term accuracy, and due to its complexity tends to operate on a road network too simplified and a zonal system too coarse to model slow modes such as walking and cycling (Cervero, 2006). SpNA requires less effort, notionally trades off short-term correlation for a longer term view, and operates at high resolution on an unsimplified network, though could benefit from increased parametrization. The current article presents a model to fill this gap. The immediate aim is to provide a high-resolution methodology for modeling cycling, but we do not preclude other uses: the long-term focus of SpNA could potentially be used to develop simplified long-term models.

Our method extends previous SpNA models calibrated to flows. In contrast to existing SpNA models, we use regularized regression tuned by cross-validation to handle multivariate models without risking overfit. This allows us: (i) to include multiple trip purposes and agglomeration effects and (ii) to quickly tune a distance decay model. Gao et al. (2013) identifies distance decay as lacking from conventional betweenness, but does not provide an efficient technique for calibrating distance decay betweenness against traffic flows. Among the 528 citations of Hillier and Iida (2005), regularized regression is not used except once on-time series kernels in Haworth (2014). We introduce the idea of weighting to optimize the Geoffrey E. Havers (GEH) statistic (commonly used as a measure of transport model quality—see Section 2.6) to replace variable transformation in a multivariate model. The combination is

effective, as the GEH weighted regularized regression is automated and therefore improves model predictions without undue extra burden on the analyst, who must only choose which betweenness computations to combine. It is therefore proposed as a methodology for practical forecasting tasks.

The model is implemented using the publicly available spatial design network analysis (sDNA+) toolbox (Cooper, Chiaradia, & Webster, 2011), which functions as either a QGIS or ArcGIS plug-in. Betweenness and link density are computed with the sDNA Integral tool, and outputs are combined and calibrated to flows using the open source sDNA Learn and Predict tools. sDNA Learn internally makes use of the glmnet package in R (Friedman, Hastie, & Tibshirani, 2009).

1.2 Modeling cycling

Despite numerous studies of cycling-mode choice (Ewing et al., 2014; Parkin, Wardman, & Page, 2007; Wardman et al., 2007; Winters, Brauer, Setton, & Teschke, 2013) and route choice (Broach, Dill, & Gliebe, 2012; Ehrgott, Wang, Raith, & van Houtte, 2012), none so far have been turned into a general purpose tool for modeling cycling in the manner of the four-step model. Such a model would have applications in estimating change to mode choice from proposed cycle infrastructure—the key economic justification for investment—as well as highlighting hotspots where new infrastructure would be useful in the first place, assisting with option selection, and illustrating how proposed infrastructure fits in to the wider network (Forsyth & Krizek, 2011; Krizek, Handy, & Forsyth, 2009). As an example of the SpNA methodology, we present a model which was used to inform production of a city-wide integrated network map, a forward plan for cycle infrastructure mandated by the Wales Active Travel Act (2013).

The lack of prior models of this type is understandable in light of the fact that motor transport models typically work at transport analysis zone (TAZ) level and thus miss small features that can influence cyclist decision making—such as the availability of residential streets mostly free from traffic, or distribution of land use within the zone that encourages shorter trips. Indeed a huge proportion of cycling trips are intrazonal rather than interzonal. Motor transport simulations also tend to exclude minor roads from the model while cyclists will make extensive use of these not only to reach endpoints but often throughout the entire trip. The answer is not simply to make smaller models, however; instead, equally wide-scale models are needed but with a large increase in resolution. There are two reasons for this. Firstly, cyclists are capable of making relatively long (city-wide) trips anyway. Secondly, the presence of motor traffic has a huge effect on cyclist behavior. If we wish to model this effect in full, an increase in resolution of the *vehicle* model is also needed to inform the cycling model. Such an increase in resolution, however, entails a greater cost in computation and calibration of the four-step model, which is typically too expensive to apply to small-scale cycling infrastructure projects in the first place.

An additional problem in applying the four-step model to cycling is the exclusion of land use-accessibility feedback effects (responsible for some notable vehicle modeling failures such as the Newbury bypass; see Atkins, 2006). These have been shown to be relevant in cycling infrastructure through mechanisms

such as residential self-selection (Cervero, 2006) in which keen cyclists will choose to live near cycling infrastructure to allow them to cycle, e.g., for the journey to work. Although extensions to the four-step model do exist which include land use, these are still considered experimental (Department for Transport, 2014c section 4.6.6, see also 2014d) and in any case are even more expensive to calibrate.

In the absence of four-step models, predictions of cycling have relied on applying an exogenous growth factor to current behavior (Schwartz et al., 1999), modeling demographics and investment in infrastructure at a course spatial scale (Parkin et al., 2007) or modeling demographics and localized spatial variables but without explicit consideration of routes traveled (Griswold, Medury, & Schneider, 2011). The latter two studies are successful at predicting mode choice ($r^2 = 0.81$) and flows ($r^2 = 0.60$), respectively, but the models are not sensitive to the precise location of infrastructure in relation to route choice and hence mode choice, and thus will ultimately be limited in their ability to suggest optimal locations for new infrastructure. An alternative approach (Lovelace et al., 2016) is to model potential rather than predictions, where potential is defined as travel demand over distances short enough to be cycled. These models are valuable for identifying potential at coarse spatial level but once that has been established, a different model is needed to predict the effect of spatially detailed infrastructure changes. Hollander (2016) argues that modeling potential, combined with mapping changes in accessibility from proposed infrastructure, is preferable to full cost-benefit modeling due to the high level of unknowns in cycle models. But what is accessibility for cyclists if not the thing that best predicts cyclist behavior? We argue that our notion of accessibility should itself be derived from a demand model. In cases where there are concerns over the risk of model failure, these can be mitigated not by eschewing modeling, but by using the best model possible to compute accessibility change and potential users of infrastructure, while stopping short of ambitious predictions of the cost-benefit ratio.

The models presented are thus high-resolution, wide-scale models of cyclist behavior based on SpNA principles. Land use-accessibility feedback is taken into account by inferring origins and destinations from the network itself. Recall from Section 1.1 the fundamental premise of SpNA; that accessibility itself shapes land use, ultimately creating origins and destinations which give rise to flows.

SpNA has been applied to cycling problems before (Cooper, 2017; Law, Sakr, & Martinez, 2014; Manum & Nordstrom, 2013; Raford, Chiaradia, & Gil, 2007) but without the nuanced microeconomic behavioral foundations used here, namely, the use of a cycling-specific distance metric both for defining network radius and route choice. Also novel in the current study is the multivariate approach, which can be interpreted as multiple betweenness computations (each of which can be seen as an agent model) combined through ridge regression to fit measured cyclist count data on the network. As well as demonstrating a distance decay model, the multivariate approach is extended to calibrate models of agglomeration economics (Bettencourt, 2013), multiple trip purposes and varying individual preferences of cyclists. The latter is acknowledged as a key issue in uptake of cycling as some cyclists are more confident in motor vehicle traffic than others.

Limitations of the cycling model are: (1) that it is currently unimodal, or at least, the effect of other modes is considered in the aggregate, with no spatial variation as would be caused, e.g., by access to a good metro system in specific locations (for simplicity Department for Transport, 2014a endorses such approaches); (2) that it does not model congestion, although in most cities this is not currently a problem for cyclists. The fact that it is a one, rather than four-step model gives the advantage of a simpler calibration process with reduced data requirements, but the disadvantage of fewer opportunities to verify the model at each step. To mitigate this construct and test separate models for mode choice data and flow data, note however that these models (unlike the four-step model) are independent so each can be used in isolation.

2. Methodology

2.1 Definition of distance

The first step to producing a behaviorally accurate cyclist model using SpNA is to determine an appropriate definition of distance through the network. The metric used is based on Cooper (2017) which chooses a subset of factors identified in the cyclist route choice study of Broach et al. (2012), informed by availability of the relevant data. Creating a model sensitive to motor vehicle traffic is considered essential, however, so a submodel is used to predict motor vehicle flows. The definition of distance applicable to cyclists is then determined by a combination of distance, straightness, slope, and motor vehicle traffic:

$$
\begin{aligned}
cyclist\ distance = {} & Euclidean\ network\ distance \times slopefac^s \\
& \times trafficfac^t + cumulative\ angular\ change \\
& \times \frac{67.2}{90} \times a
\end{aligned} \tag{1}
$$

$$
slopefac = \begin{cases} 1.000 & if\ slope < 2\ \% \\ 1.371 & if\ 2\ \% < slope < 4\ \% \\ 2.203 & if\ 4\ \% < slope < 6\ \% \\ 4.239 & if\ slope > 6\ \% \end{cases} \tag{2}
$$

$$
trafficfac = 0.84\ e^{\frac{AADT}{1000}} \tag{3}
$$

in which AADT is annual average daily (vehicle) traffic. This relationship is taken to be applicable to any path through the network on which slope and AADT remain constant—if these conditions hold, it does not matter whether Eq. (1) is applied to links, origin-destination paths or junctions. In reality, slope and AADT vary, so the unit of computation matters. sDNA (set to "discrete space" mode) will compute distance between neighboring links as the sum of: (1) distance from first link center to junction; (2) distance through junction (comprising only the contribution from angular change); and (3) distance from junction to center of second link. Longer distances through the network are taken as the sum of distances between successive links on the path. Thus, influence of slope and AADT is computed separately for each half link.

The structure and constants of Eqs. (1)–(3) are chosen to match Broach et al., leaving room for calibration by changing s,

t, and a; with the exponential form of Eq. (3) derived by fitting a curve to Broach's fixed distance bands in order to achieve better control over calibration. To match the original study we would set

$$a = 1$$
$$s = 1 \qquad . \qquad (4)$$
$$t = 0.05$$

Cooper (2017) found the following parameter values best fit the Cardiff data in a homogenous model:

$$a = 0.2$$
$$s = 2 \qquad . \qquad (5)$$
$$t = 0.04$$

We take these as a starting point for the current study, but later introduce heterogeneous models in which different agents have different values of t, the aversion to motor traffic.

All cyclist distances are measured as round trip distances using the same route for the outward and return journey, as a cyclist who goes downhill knows they must later climb back up again, and this will affect their decision to cycle.

2.2 Definition of betweenness

Having defined distance appropriately, we apply it to the SpNA concept of link weighted betweenness. This is a flow model in which, for each pair of link centroids on the network, a single agent is presumed to travel by the shortest path between them. We keep count on each link, of the number of agents which pass along it. This can be stated (for networks with unique shortest paths) as

$$
\text{Betweenness}\,(x, r, d_{routing}, d_{radius}) \\
= \sum_{y \in N} \sum_{z \in R(y, r, d_{radius})} OD(y, z, x, d_{routing}), \qquad (6)
$$

$$
OD(y, z, x, d_{routing}) = \\
\begin{cases}
1 & \text{if } x \text{ is on the shortest path from } y \text{ to } z \text{ as defined by metric } d_{routing} \\
1/2 & \text{if } x = y \neq z \text{ or } x = z \neq y \\
1/3 & \text{if } x = y = z \\
0 & \text{otherwise}
\end{cases}
\qquad (7)
$$

where x, y, and z are links in the network N. $R(y, r, d_{radius})$ is the subset of the network closer to link y than a threshold radius r defined according to distance metric d_{radius}. The OD (y, z, x, d) function defined in Eq. (7) describes the proportion of link x that falls on the shortest path from the middle of link y to the middle of link z, with partial contributions for links which form the endpoints of the shortest path. This is the definition used for univariate models in the current study. For multivariate models, to reduce multicollinearity we choose to introduce both a minimum and maximum threshold radius to

the formula, hence in our case,

$$
\text{Betweenness}\,(x, rmin, rmax, d_{routing}, d_{radius}) \\
= \sum_{y \in N} \sum_{z \in R(y, rmin, rmax, d_{radius})} OD(y, z, x, d_{routing}). \qquad (8)
$$

Where $R(y, rmin, rmax, d_{radius})$ is the subset of the network closer to link y than a threshold radius $rmax$ but further from y than $rmin$. SpNA literature often uses different metrics for $d_{routing}$ and d_{radius} (see Cooper, 2015 for a discussion of implications) though in the current case we define both according to the *cyclist distance* concept above, except for the baseline model which uses Euclidean network distance for d_{radius}.

2.3 Distance decay model for flows and mode choice

Having defined distance and betweenness, we compute betweenness within a number of cycling distance bands. In all but the simplest model, we define both $d_{routing}$ and d_{radius} in terms of cyclist distance, therefore the same distance factors (Euclidean network distance, slope, straightness, and vehicle traffic flow as combined in Eq. 1) are taken to influence both mode (the decision to cycle) and route choice. The distance bands chosen for cyclists are round trip distances of 3, 5, 8, 11, 15, and 20 cyclist-adjusted km. Multiband betweenness is then combined using multivariate regression to fit link flow data, such that

$$
\text{flow on link} = \beta_1 Bt_1 + \beta_2 Bt_2 + \cdots, \qquad (9)
$$

where Bt_1, Bt_2 are the betweenness in distance bands 1, 2, etc.; and the βs are regression coefficients. The process is illustrated in Figure 1. Note that fitting this model constitutes nonparametric fitting of a distance decay curve by regression. The high level of detail in spatial network models means they are more computationally demanding than mainstream vehicle assignment models, however, reducing the calibration process to a linear regression problem in this manner keeps computation times manageable, and means that subsequent models can often reuse the same betweenness computation outputs.

The cycling-mode choice model is independent from the flow model, but is also based on distance decay. Instead of computing betweenness, the variables used to predict mode choice are simple counts of network quantity (measured by number of links) within each distance band. The process is shown in Figure 2. Together, the distance bands form a multidimensional definition of accessibility, in which both quantity and quality of access to destinations is captured.

The submodel used to compute vehicle flows to inform the cyclist model is similar. The differences are: (1) $d_{routing}$ is set to cumulative angular change alone, which in an urban environment is a good approximation to vehicle route choice, often superior to minimizing actual travel time (Ciscal-Terry, Dell'Amico, Hadjidimitriou, & Iori, 2016; Jayasinghe, 2017, chapter 3). (2) d_{radius} is defined in network Euclidean distance. (3) The radius used is a single maximum one-way trip distance, calibrated to fit the available count data.
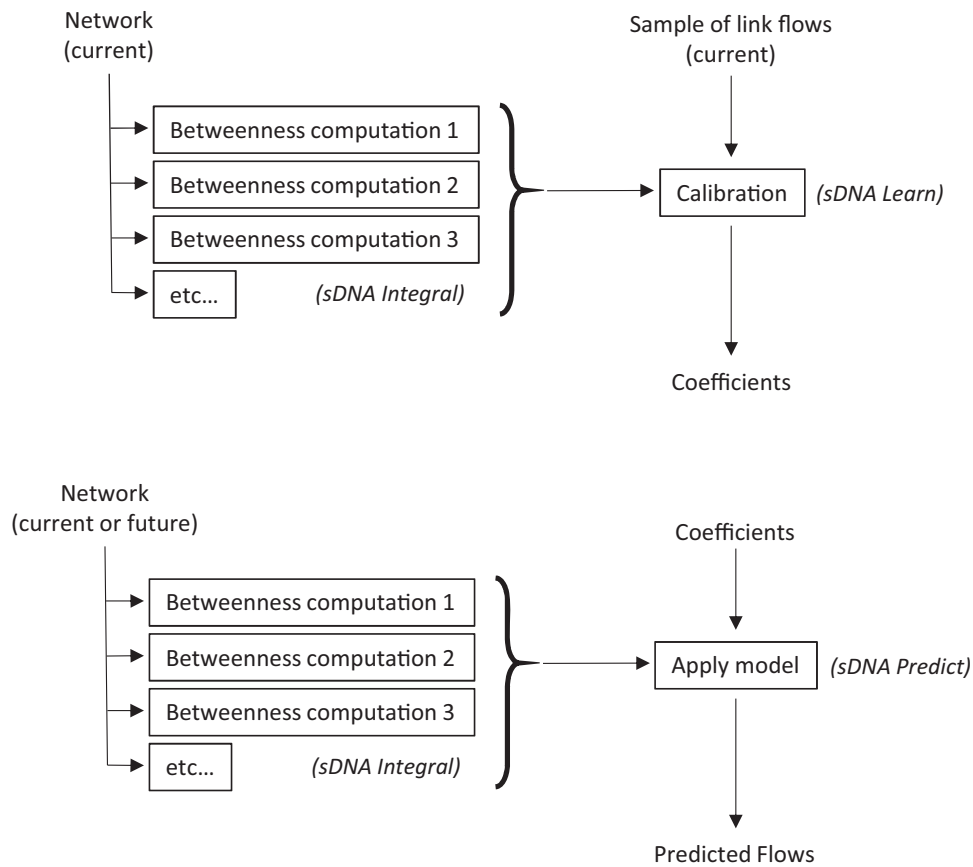
Figure 1. Data flow diagram for the flow model.

2.4 *Further models for cyclist flow*

A key aspect of the methodology which generalizes beyond models of cycling, is that multiple predictor variables can be constructed from multiple betweenness computations, and then combined through regression to fit observed behavior. To demonstrate this in the current case, we present three further variables to improve prediction of cyclist flows compared to the distance decay model defined in Section 2.3 above.

2.4.1 *Agglomeration economics*

Although the definition of betweenness already implies some degree of land use-accessibility feedback, it is preferable to calibrate this effect more explicitly. It is well known that more accessible land (e.g., city centers) exhibits denser land use, and there is a nonlinear scaling effect for economic activity in progressively larger cities (Bettencourt, 2013). To calibrate such a model efficiently, we add extra variables to the distance decay model that relate to flows at the same distance bands as before, but *only to and from the densest 30% of network links*. Density is defined as number of links within a 2-km Euclidean network radius. Both of these figures (30%, 2 km) are chosen manually to give an approximate match to what we perceive to be the city center. Thus, analogously to the distance decay model, we have now created a two-band model of urban density which is fitted nonparametrically as before.

2.4.2 *Heterogeneous agent traffic aversion*

The interpretation thus far presented is that the distance decay and agglomeration models can be seen as using regression to perform nonparametric calibration of distance decay and economic scaling curves. However, these models can also be interpreted as incorporating different types of behavior: trips of varying lengths from everywhere to everywhere plus trips of varying lengths to the city center. The regression model thus chooses an appropriate balance of behaviors that best explains the observed flows. Two more variables are presented which further extend this behavioral concept. The traffic aversion variables achieve this by adding betweenness computations based on different levels of aversion to traffic. Referring to Eq. (1), we compute betweenness with $t = 0.06$ and $t = 0.08$ as well as $t = 0.04$.

2.4.3 *Introducing agents on purely recreational trips*

Choice of the final variable is informed by inspection of residuals from the combination previous variables. In the case of Cardiff, the residuals reveal that flow on the Taff Trail and connecting routes is underpredicted. The Taff Trail is a flagship cycle route along the river of the same name, which flows through a green corridor formed by a near-continuous series of parks crossing the entire city in a north-south direction. It is hypothesized that the greater-than-expected use of this facility is caused by recreational use of the trail as an end in its own right; therefore extra variables are included to model behavior
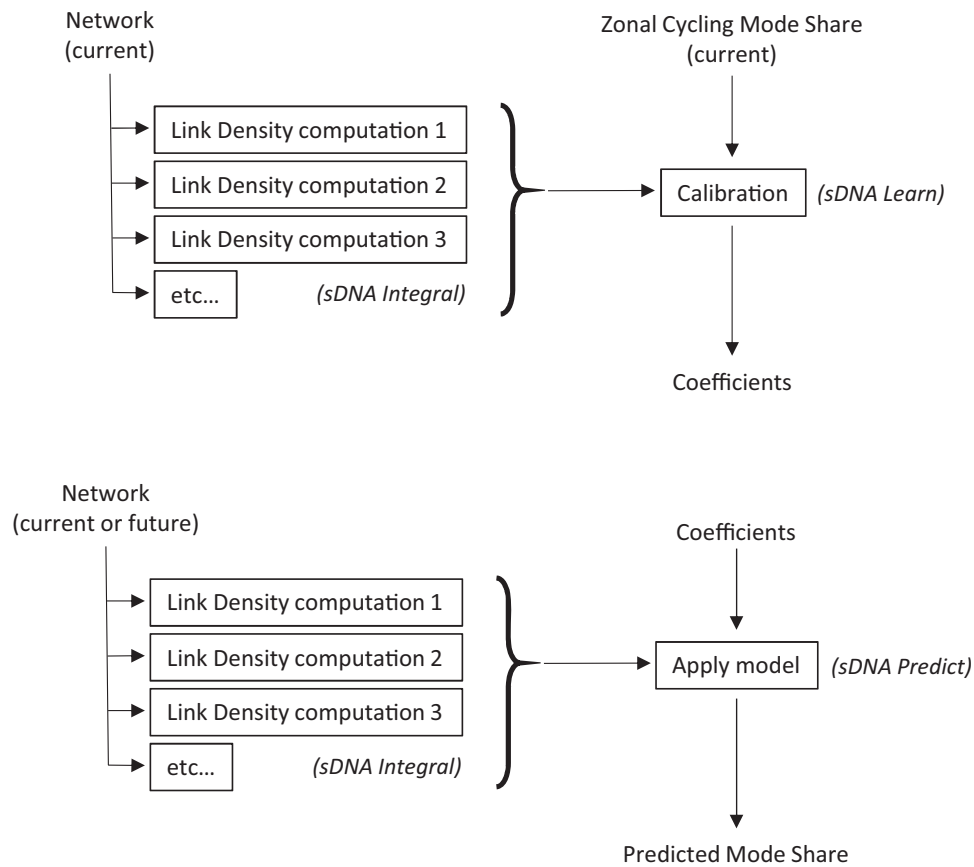
Figure 2. Data flow diagram for the mode choice model.

of cyclists going to the trail itself. This is again a betweenness computation with the Taff Trail as a destination; this time weighted by destination trail length.

2.5 Data sources

Road network data are based on OpenStreetMap (2015), which at time of writing contains more information on traffic-free cycle routes than any other publicly available routable network data (including commercial offerings, Lovelace, 2015). The network data are prepared according to the instructions in the sDNA user manual, including planarization and use of a high cluster tolerance to correct errors (Cooper, 2016).

Flow models are calibrated to measurements of cycle flows for 107 locations on roads (Department for Transport, 2014e) and 14 locations on traffic-free paths (provided by Cardiff Council). Traffic-free path counts are derived from electronic counters covering a 3-month period plus a year-round counter which is used to deduce a scaling factor to estimate average annual daily count. This differs from the on-road counting methodology (Department for Transport, 2011) which is likely to undercount cyclists. In the final model therefore, a dummy variable is introduced to account for data source, to estimate the effect of differing count methodology between the data sets.

Mode choice models are calibrated to journey-to-work data (proportion of working population travelling by bicycle) for 1,077 census output areas covering Cardiff (Office for National Statistics, 2011).

2.6 Regression techniques

Baseline univariate models are fitted using ordinary least squares linear regression. In univariate models, both predictor and target variables are Box-Cox transformed to reduce the influence of outliers, trade-off relative versus absolute error, and provide comparability with previous work that uses power transforms.

All multivariate models are fitted using ridge regression (Amemiya, 1985; Tikhonov, 1943) in order to correctly handle the multicollinearity inherent in different betweenness computations on the same network. Coefficients are constrained to be positive and no variable transform is used. The ridge regression can be interpreted in multiple ways. From a frequentist perspective, it applies a penalty to coefficient sizes in linear regression to prevent overfit. From a Bayesian perspective, it is applying a prior distribution on the coefficients with zero being the most likely value; i.e., a variable is assumed to have no effect unless we can strongly demonstrate otherwise. From the perspective of transport modeling history, this can also be considered as a mild form of entropy maximization.

All model performance is reported using generalized cross-validation, with seven folds and 50 bootstrap repetitions to achieve stable estimates both of cross-validated performance and (in multivariate models) of the optimal ridge penalty parameter. This achieves dual aims: (1) as data collection is expensive we make maximum use of the available data for calibration, but (2) r^2 values report the success of model predictions on test data, not just model fit. This is crucial to ensure

that the more complex models presented are genuinely useful at out-of-set prediction rather than data-dredged or overfitted.

As is traditional in transport models we report (with the aim of minimizing) the GEH statistic,

$$GEH = \sqrt{\frac{2(prediction - count)^2}{prediction + count}} \qquad (10)$$

where *prediction* and *count* are measured in units per hour, and GEH hence has units of $\sqrt{units\ per\ hour}$. Typically a vehicle model is considered usable if GEH < 5 for 85% of measured link flows (Department for Transport, 2014b, section 3.2.7). However, as GEH is not scale free, this target is too easily obtained for cycling models where flow counts are much smaller than vehicle flow counts. Rather than attempting to determine an appropriate GEH target for cycle models, we take an approach that is free of scaling effects.

The spirit of GEH is that it strikes a compromise between the reporting of relative and absolute errors, both of which are considered important to minimize. Note that while absolute errors are minimized through use of an ordinary least squares fit, relative errors can be minimized by log transformation of the data before regression, and a balance of both can be achieved using a Box-Cox (1964) transform. SpNA has previously used this (Cooper, 2015; Cooper & Chiaradia, 2015) or the simpler cube root transform approach (Hillier & Iida, 2005; Turner, 2007). However, in multivariate analysis such transformation is undesirable as altering the model structure such that that y is not a linear sum of the xs violates the physical interpretation of link flow as being the sum total of multiple individual agent behaviors. We instead replicate the absolute/relative error trade-off determined by the Box-Cox power parameter λ (not to be confused with the Tikhonov regularization parameter of the same name) by weighting each data point y with the ratio between its transformed and untransformed value:

$$weight(y, \lambda) = \frac{y^\lambda}{y}. \qquad (11)$$

Practical experimentation shows that values of around $\lambda = 0.7$ minimize the unweighted average GEH and thus represent the same relative prioritization of absolute and relative error. Model performance is thus reported as weighted r^2, which avoids the scale-dependency problems of GEH while also reflecting the same balance of priorities.

3. Results

The vehicle traffic submodel showed optimum fit to the data with $r^2 = 0.81$ for a 28-km (one-way) trip distance.

Table 1 shows results for the cycling flow models. Compared to the baseline model (univariate Euclidean radius with cyclist route choice: $r^2 = 0.56$), use of hybrid radius (i.e., defining mode as well as route choice in terms of cyclist distance) substantially improves model fit to 0.63. Replacing the Box-Cox transform with the GEH weighting scheme reduces model performance to 0.42; this is to be expected (1) as nonlinearities in response are no longer captured and (2) as changing the

Table 1. Cross-validated fit for cyclist flow models.

Flow Model	Transform & weighting	r^2, cross-validated
Univariate, calibrated Euclidean radius (6 km round trip)	Box-Cox; no weighting	0.56
Univariate, calibrated Hybrid Radius (8 km)		0.63
Univariate, calibrated Hybrid Radius (8 km)	No transform; weighting as GEH	0.42
Multiple radius, medium traffic aversion only (t = 0.04)		0.65
Multiple radius, agglomeration effects, trips to center and recreation, medium traffic aversion only (t = 0.04)		0.73
As above with mixed traffic aversion (t = 0.04, 0.06, 0.08)		0.75
As above with dummy variable accounting for data source		0.78

weighting means that r^2 now reports a model fit for the more difficult problem of optimizing GEH. However, this allows correct specification of multivariate models, which more than compensate for the performance drop, with the best of them (multiple radius, agglomeration effects, recreation trips, mixed traffic aversion, and correcting for data source) giving $r^2 = 0.78$.

The best model has GEH < 5 for 93% of data points, versus 88% for the baseline model. Mean GEH is 1.9 and 2.3, respectively. GEH is computed for the peak hour based on estimation that the peak carries 10% of daily flow.

Table 2 shows results for mode choice. The baseline model based on links in Euclidean radius only (i.e., urban density) has no clear peak for calibrating the appropriate radius, but gives correlation of 0.41. Defining urban density in terms of cycling distance increases correlation to 0.43 with a clear peak for a 13-km (cyclist-adjusted) round trip. Use of a multivariate model gives the best correlation at 0.45. Note that numerous sociodemographic and promotional factors are known to affect mode choice (Parkin et al., 2007) so we should not expect to explain much more of the variance from network design alone.

An example of predicted cyclist flow is shown in Figure 3. The ability of the model to provide detailed spatial information on the links between infrastructure and cycling potential is illustrated in Figure 4, which shows potential increase in mode choice in the hypothetical situation that all routes were free of motor traffic (the zero-traffic scenario). Figure 5 shows increase in flow under the same scenario. While the scenario may not be attainable in practice, the same information can be interpreted as an accessibility model which highlights trip endpoints

Table 2. Cross-validated fit for mode choice models.

| Mode choice model	Transform (unweighted)	r^2, cross-validated
Univariate, calibrated Euclidean radius (urban density only)	Box-Cox predictor and target	0.41
Univariate, calibrated hybrid radius (urban density adjusted for cyclist distance)		0.43
Multivariate, hybrid radius	Box-Cox target	0.45
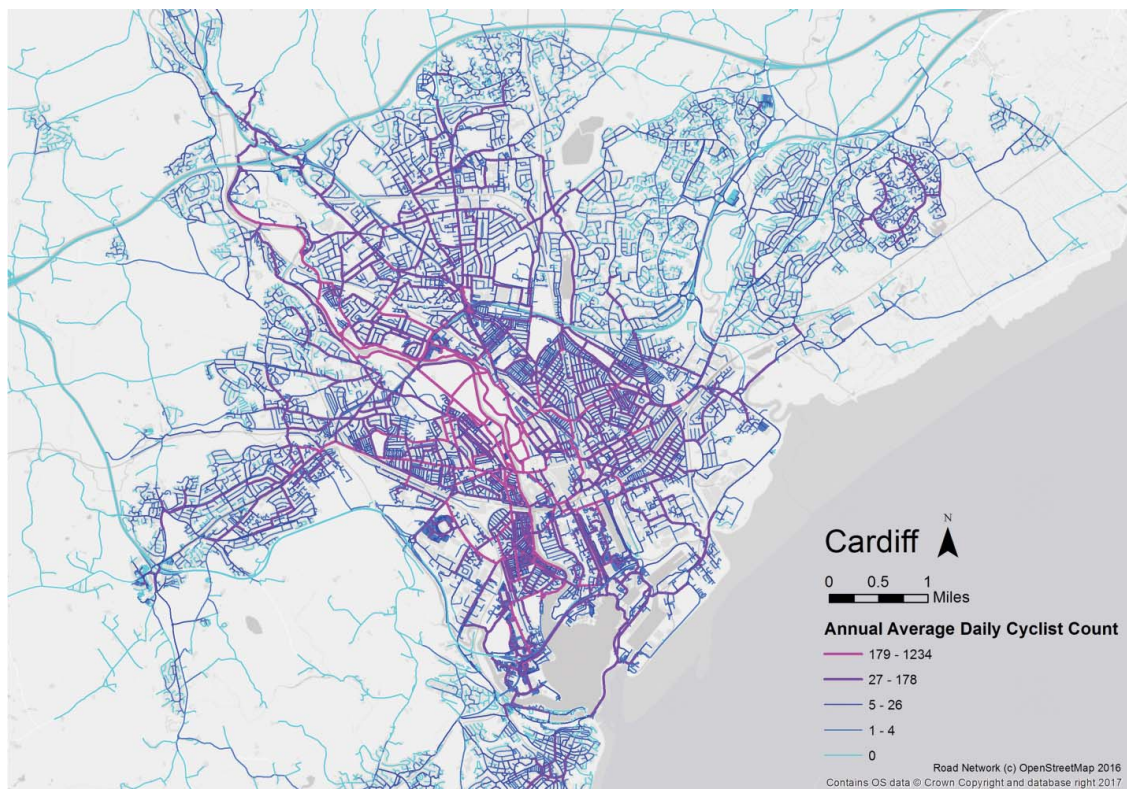
Figure 3. Predicted cyclist flows.
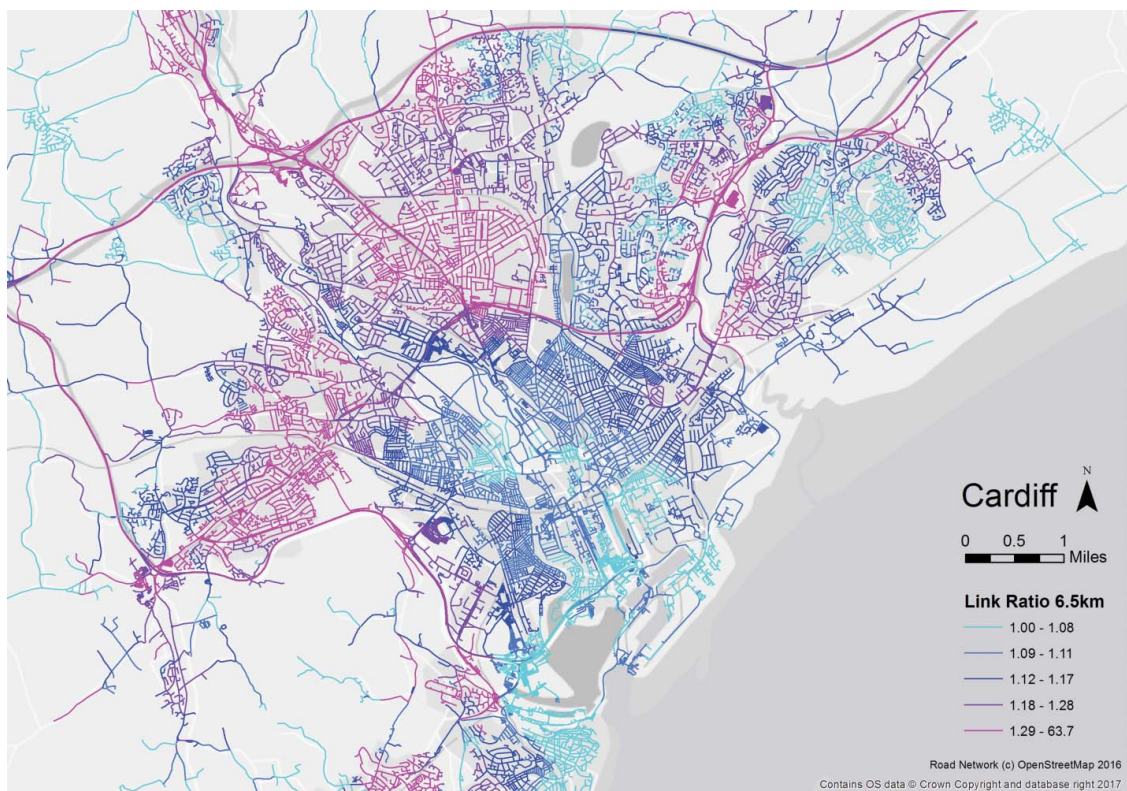
Figure 4. Potential for improvement by introducing traffic-free routes. Link ratio shows increase in network links accessible within 6.5 km of "cyclist distance" in the zero-traffic scenario; this is strongly correlated with cycling mode choice, hence high values of link ratio show greater potential for increasing levels of cycling through improved accessibility. This model suggests that potential for modal shift is higher in the suburbs than city center as the centre already benefits from compact urban form; however, Figure 5 shows that it is necessary to build infrastructure in both places, and also linking the two together, in order to unlock the potential increase.
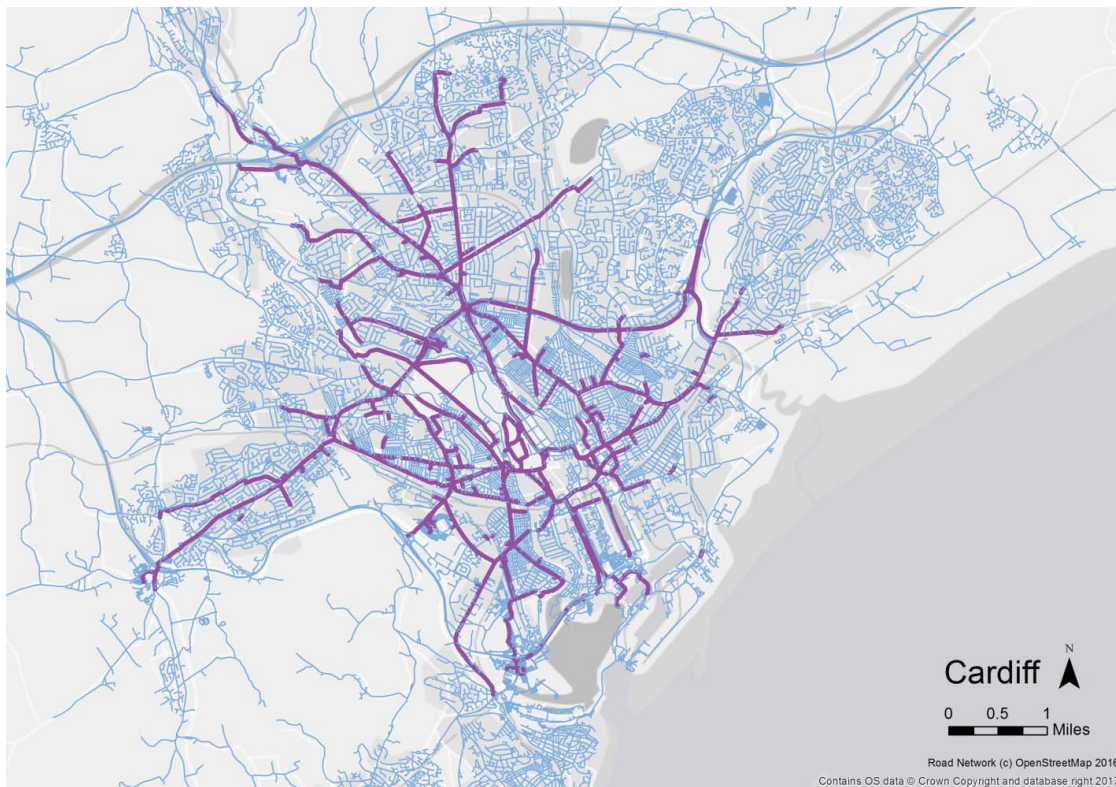
Figure 5. Routes that would be more heavily used in absence of vehicle traffic, and thus suggest fruitful locations for infrastructure investment. These are outliers when subtracting baseline flow from flow in the zero-traffic scenario.

(Figure 4) and routes (Figure 5) currently worst affected by motor vehicle traffic. In particular, routes that would be more heavily used in the absence of motorized traffic suggest locations for investment in infrastructure that improves both perceived and actual cyclist safety; this could be provided either on the route in question, or a suitable parallel route.

These visualizations have application in targeting infrastructure improvement schemes prior to the design phase. Once specific options are designed they can be simulated by the same model, and flow filtering techniques (constraining the model to show only flows through the new infrastructure) can help visualize interaction of proposed infrastructure with the wider network.

4. Conclusions

This article has presented a methodology for spatially detailed transportation modeling on a city-wide scale, based on SpNA and cross-validated ridge regression, and without need for origin or destination data. Although many factors considered in four-step modeling are excluded, this is a step forward in terms of scale, as four-step models do not typically include every link in a city. It also represents a step forward for SpNA methodologies which do not normally include distance decay, cross-validation, and a reliable method of combining multiple betweenness calculations without overfit; nor are they optimized for performance in terms of the GEH statistic popular in transport planning.

The methodology has been demonstrated on models of cycling; however, it is also applicable to other modes:

pedestrian, vehicle, public transport, and even multimodal networks. At its core is combination of multiple SpNA variables which a practitioner can choose from depending on their own problem constraints and available time. This gives the expressiveness of a heterogeneous agent model albeit with easier calibration and standardized behavior based on well-known network analysis measures. There are likewise a multitude of ways in which model outputs can be presented, including accessibility maps, to inform spatially sensitive models of potential for behavior change (in this case, potential for cycling).

While the mathematics may be unfamiliar to transport practitioners, the practical modeling process is relatively simple and inexpensive in terms of data collection. The sDNA Integral software is run a number of times to compute different agent behaviors, which are then calibrated against real data with sDNA Learn and extrapolated with sDNA Predict. The software can also interface with existing models through export of skim matrices to model accessibility at high resolution, and import of OD matrices to use in the assignment phase (Cooper, 2016).

Compared to spatial network modeling traditions, it is noted that removing the nonlinear Box-Cox transform from the regression results in a decrease in model performance, in part because nonlinear effects are excluded. Although performance is subsequently regained through multivariate modeling, it might thus be expected that nonlinear multivariate penalized regression could increase performance even further. However, this would in turn be a poor substitute for full parameterization of demand elasticity, which is the likely source of these

nonlinearities. This seems a likely future direction for SpNA models to take, particularly if we aspire to better representation of land use-transport effects. The disadvantage will be a large increase in computing resources needed, as it may not be possible to fully reduce a variable elasticity model to a linear regression problem as was achieved for distance decay and agglomeration effects in this study.

Finally, we have shown that (at least in the United Kingdom) existing targets for the GEH statistic are perhaps too easily achieved with the small flow numbers present in cycling models. While we used the scale-free r^2 statistic to evaluate overall model performance, the best choice of metric to describe fit of individual data points remains an open question.

Acknowledgments

ORCID

Crispin H. V. Cooper 🄳 http://orcid.org/0000-0002-6371-3388

References

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Atkins, (2006). *A34 Newbury Bypass – 5 Years After, Evaluation*. Retrieved from http://webarchive.nationalarchives.gov.uk/20120810121037/http://www.highways.gov.uk/roads/documents/Newbury_Bypass_Five_Years_After_1.pdf

Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science, 340* (6139), 1438–1441. doi:10.1126/science.1235823.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211–252. Retrieved from http://www.jstor.org/stable/2984418

Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice, 46*(10), 1730–1740. doi:10.1016/j.tra.2012.07.005.

Cervero, R. (2006). Alternative approaches to modeling the travel-demand impacts of smart growth. *Journal of the American Planning Association, 72*(3), 285–295. doi:10.1080/01944360608976751.

Chiaradia, A. J., Cooper, C. H. V., & Wedderburn, M. (2014). Network geography and accessibility. In *Proceedings of 12th Transport Practitioners' Meeting*. London, PTRC.

Ciscal-Terry, W., Dell'Amico, M., Hadjidimitriou, N. S., & Iori, M. (2016). An analysis of drivers route choice behaviour using GPS data and optimal alternatives. *Journal of Transport Geography, 51*(Supplement C), 119–129. doi:10.1016/j.jtrangeo.2015.12.003.

Cooper, C. H. V. (2015). Spatial localization of closeness and betweenness measures: A self-contradictory but useful form of network analysis. *International Journal of Geographical Information Science, 29*(8), 1293–1309. doi:10.1080/13658816.2015.1018834.

Cooper, C. H. V. (2016). Spatial design network analysis (sDNA) version 3.4 Manual. Retrieved 15 September 2016, from http://www.cardiff.ac.uk/sdna/software/documentation

Cooper, C. H. V. (2017). Using spatial network analysis to model pedal cycle flows, risk and mode choice. *Journal of Transport Geography, 58*, 157–165. doi:10.1016/j.jtrangeo.2016.12.003.

Cooper, C. H. V., & Chiaradia, A. J. (2015). sDNA: How and why we reinvented Spatial Network Analysis for health, economics and active modes of transport. In Nick Malleson (Ed.), *GISRUK 2015 Proceedings*. Leeds. doi:10.6084/m9.figshare.1491375.

Cooper, C. H. V., Chiaradia, A. J., & Webster, C. (2011). Spatial Design Network Analysis (sDNA). Retrieved 15 September 2016, from www.cardiff.ac.uk/sdna

Cooper, C. H. V., Fone, D. L., & Chiaradia, A. (2014). Measuring the impact of spatial network layout on community social cohesion: A cross-sectional study. *International Journal of Health Geographics, 13* (1), 11. doi:10.1186/1476-072X-13-11.

Department for Transport. (2011). *Road traffic estimates methodology note*, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230528/annual-methodology-note.pdf

Department for Transport. (2014a). *Active mode appraisal* (Transport Analysis Guidance No. A5.1), UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/370544/webtag-tag-unit-a5-1-active-mode-appraisal.pdf

Department for Transport. (2014b). *Highway assignment modelling* (Transport Analysis Guidance No. M3.1). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427124/webtag-tag-unit-m3-1-highway-assignment-modelling.pdf

Department for Transport. (2014c). *Principles of modelling and forecasting* (Transport Analysis Guidance No. M1). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427118/webtag-tag-unit-m1-1-principles-of-modelling-and-forecasting.pdf

Department for Transport. (2014d). *Supplementary guidance: Land use/transport interaction models* (Transport Analysis Guidance). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427140/webtag-tag-supplementary-luti-models.pdf

Department for Transport. (2014e). Traffic Count Data. Retrieved from http://www.dft.gov.uk/traffic-counts/

Ehrgott, M., Wang, J. Y. T., Raith, A., & van Houtte, C. (2012). A bi-objective cyclist route choice model. *Transportation Research Part A: Policy and Practice, 46*(4), 652–663. doi:10.1016/j.tra.2011.11.015.

Ewing, R., Tian, G., Goates, J. P., Zhang, M., Greenwald, M. J., Joyce, A., … Greene, W. (2014). Varying influences of the built environment on household travel in 15 diverse regions of the United States. *Urban Studies, 52*(13), 2330–2348. doi:10.1177/0042098014560991.

Fone, D., Dunstan, F., White, J., Webster, C., Rodgers, S., Lee, S., … Lyons, R. (2012). Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). *BMC Public Health, 12*(1), 428. doi:10.1186/1471-2458-12-428.

Forsyth, A., & Krizek, K. (2011). Urban design: Is there a distinctive view from the bicycle? *Journal of Urban Design, 16*(4), 531–549. doi:10.1080/13574809.2011.586239.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41. doi:10.2307/3033543.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning and Design, 40*(1), 135–153. doi:10.1068/b38141.

Griswold, J. B., Medury, A., & Schneider, R. J. (2011). Pilot models for estimating bicycle intersection volumes. *Safe Transportation Research & Education Center*. Retrieved from http://escholarship.org/uc/item/380855q6

Haworth, J. (2014). *Spatio-temporal forecasting of network data*. UCL (University College London).

Hillier, B., & Iida, S. (2005). Network and psychological effects in Urban movement. In A. G. Cohn & D. M. Mark (Eds.), *Spatial information theory* (pp. 475–490). Berlin, Heidelberg: Springer. Retrieved from http://link.springer.com/chapter/10.1007/11556114_30

Hollander, Y. (2016, January). Justifying investments in cycling infrastructure: 10 lessons learnt. Retrieved from http://www.ctthink.com/publications.html

Jayasinghe, A. B. (2017). *A network centrality-based simulation approach to model traffic volume* (Ph.D. Thesis). Nagaoka University of Technology.

Krizek, K. J., Handy, S. L., & Forsyth, A. (2009). Explaining changes in walking and bicycling behavior: Challenges for transportation research. *Environment and Planning B: Planning and Design*, 36(4), 725–740. doi:10.1068/b34023.

Law, S., Sakr, F. L., & Martinez, M. (2014). Measuring the changes in aggregate cycling patterns between 2003 and 2012 from a space syntax perspective. *Behavioral Sciences*, 4(3), 278–300. doi:10.3390/bs4030278.

Lovelace, R. (2015). Crowd sourced vs centralised data for transport planning: A case study of bicycle path data in the UK. In *GIS Research UK (GISRUK)*. Leeds. Retrieved from http://leeds.gisruk.org/abstracts/GISRUK2015_submission_71.pdf

Lovelace, R., Goodman, A., Aldred, R., Berkoff, N., Abbas, A., & Woodcock, J. (2016). The propensity to cycle tool: An open source online system for sustainable transport planning. *Journal of Transport and Land Use*, 10(1), 505–528. doi:10.5198/jtlu.2016.862.

Lowry, M. (2014). Spatial interpolation of traffic counts based on origin–destination centrality. *Journal of Transport Geography*, 36, 98–105. doi:10.1016/j.jtrangeo.2014.03.007.

Manum, B., & Nordstrom, T. (2013). Integrating bicycle network analysis in urban design: Improving bikeability in Trondheim by combining space syntax and GIS-methods using the place syntax tool. In *Proceedings of the Ninth International Space Syntax Symposium*. Seoul, Sejong University.

Office for National Statistics. (2011). *Method of travel to work* (No. QS701EW). Retrieved from https://www.nomisweb.co.uk/census/2011/qs701ew

Omer, I., Gitelman, V., Rofe, Y., Lerman, Y., Kaplan, N., & Doveh, E. (2017). Evaluating crash risk in urban areas based on vehicle and pedestrian modeling. *Geographical Analysis*. doi:10.1111/gean.12128.

OpenStreetMap contributors. (2015). Open Street Map.

Ortúzar, J. de D., & Willumsen, L. G. (2011). *Modelling transport* (4th ed.). Chichester, West Sussex, United Kingdom: Wiley-Blackwell.

Parkin, J., Wardman, M., & Page, M. (2007). Estimation of the determinants of bicycle mode share for the journey to work using census data. *Transportation*, 35(1), 93–109. doi:10.1007/s11116-007-9137-5.

Patterson, J. L. (2016). Traffic modelling in cities—Validation of space syntax at an urban scale. *Indoor and Built Environment*, 25(7), 1163–1178. doi:10.1177/1420326X16657675.

Raford, N., Chiaradia, A., & Gil, J. (2007). Space syntax: The role of urban form in cyclist route choice in Central London. In *TRB (Transportation Research Record) 86th Annual Meeting Compendium of Papers CD-ROM* (pp. 07–2738). Washington, DC Transportation Research Board. Retrieved from http://escholarship.org/uc/item/8qz8m4fz

Sarkar, C., Gallacher, J., & Webster, C. (2013). Urban built environment configuration and psychological distress in older men: Results from the Caerphilly study. *BMC Public Health*, 13(1), 695. doi:10.1186/1471-2458-13-695.

Sarkar, C., Webster, C., & Gallacher, J. (2014). *Healthy cities: Public health through urban planning*. Cheltenham, UK: Edward Elgar Publishing.

Sarkar, C., Webster, C., Pryor, M., Tang, D., Melbourne, S., Zhang, X., & Jianzheng, L. (2015). Exploring associations between urban green, street design and walking: Results from the Greater London boroughs. *Landscape and Urban Planning*, 143(Supplement C), 112–125. doi:10.1016/j.landurbplan.2015.06.013.

Schwartz, W. L., Porter, C. D., Payne, G. C., Suhrbier, J. H., Moe, P. C., & Wilkinson III, W. L. (1999). *Guidebook on methods to estimate non-motorized travel: Supporting documentation*. Retrieved from http://trid.trb.org/view.aspx?id = 503335

Serra, M., & Hillier, B. (2017). Spatial configuration and vehicular movement. In *Proceedings of the 11th Space Syntax Symposium*. Lisbon.

Tikhonov, A. N. (1943). Об устойчивости обратных задач. *Doklady Akademii Nauk SSSR*, 39(5), 195–198.

Turner, A. (2007). From axial to road-centre lines: A new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design*, 34(3), 539–555. Retrieved from http://eprints.ucl.ac.uk/2092 doi:10.1068/b32067.

Wardman, M., Tight, M., & Page, M. (2007). Factors influencing the propensity to cycle to work. *Transportation Research Part A: Policy and Practice*, 41(4), 339–350. doi:10.1016/j.tra.2006.09.011.

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2013). Mapping bikeability: A spatial tool to support sustainable travel. *Environment and Planning B: Planning and Design*, 40(5), 865–883. doi:10.1068/b38185.