Do road-traffic injuries scale non-linearly with travel?

Rob Johnson | https://github.com/robj411/safety_in_numbers_power_law

September 5, 2019

1 Introduction

The term "safety in numbers" reflects the observation that a change in the number of road users is not met by a linear change in the number of injuries caused to or by the road-user group. The number of injuries, I, is generally formulated as follows, in terms of a base rate, α, the number of road users of one type (say, motorists), M, the number of road users of the second type (say, cyclists), C, and "safety in numbers" exponents for each, β_1 and β_2 (Elvik and Bjørnskau, 2017):

$$I = \alpha M^{\beta_1} C^{\beta_2}. \tag{1}$$

For completeness, one might include other covariates, e.g.

$$I = \alpha M^{\beta_1} C^{\beta_2} \exp\left(\sum_{i=3}^{P} \beta_i X_i\right), \tag{2}$$

but these are not central to the present discussion, which focuses around the variables in Equation 1.

Beyond observational inference, such analyses are proposed to inform public-health forecasting (Schepers and Heinen, 2013). This includes assessment of likely health benefits following policy change, as well as forecasting healthcare needs given the expected change to transport-related behaviour, e.g. the increase in motor vehicle ownership expected in cities in fast-growing economies in the coming years. We have previously used this formulation to make predictions of numbers of road injuries for specific cities (Accra, Sao Paulo), as well as in the developing generic software ITHIM-R. (ITHIM: integrated transport and health impact model.)

In this work, we question the validity of the assumptions implicit in the model as well as its applications. Our intention here is to examine (1) the interpretation of (and language around) Equation 1, (2) its proposed application to health-impact modelling, and (3) what "safety in numbers" is and how it might be measured. We might consider two types of study: (a) small scale, and (b) inter city. In this work we focus on the latter, but make reference to the former, as it provides some insights, it might be used for prediction, and in the end we would like to have a single comprehensive framework. The picture of questions is shown in Table 1.

Using a theoretical approach, we derive the result that coefficients $\beta_1 = \beta_2 = 0.5$ learnt from multiple settings correspond to linear scaling across time and space. Put another way, these coefficients permit "tiling" of a small space to create a larger space whose properties are the same. We confirm this result through simulation, which shows also that the coefficients corresponding to variation in density given fixed space are $\beta_1 = \beta_2 = 1$. These provide null hypotheses for studies assessing the impact of road-user number on collision risk.

We present an alternative model which includes the components of Equation 1 and additionally it explicitly includes size. We use exponents δ_1 and δ_2 to parametrise this model, to distinguish it from existing models. The cases discussed above (scaling size, constant density ($\beta_1 = \beta_2 = 0.5$) and

Table 1: Questions & answers for different study types

Question	*Study level* (a) Small scale	(b) Inter city
(1) What does it mean?	(Section 2.3)	$\beta_1 + \beta_2 = 1$ means linearity (Section 3.1)
(2) How do we use it to predict?	Multiply then sum (Section 4.1)	Account for density (Sections 5 and 6)
(3) How do we study it?	(Section 2.3)	Account for density (Section 6 and Appendix C)

constant size, scaling density ($\beta_1 = \beta_2 = 1$)) are special cases of this model. We apply this model to data for England. Our results are consistent with the preceding theoretical and simulation analyses. With this model and these data, we reject the null hypothesis of linearity of injuries with respect to road-user density for most subsets of the data.

The main purpose of this correspondence is to open new avenues for research into road-traffic injury dynamics. We highlight some important features of existing methods and propose some amendments to avoid some problems, but we are unable to answer the questions set out in Table 1. We aim to highlight the areas we identify as most in need of attention, namely: accounting for size/density; the definition of "density"; whether the type of model we use is appropriate for any of our objectives; the link between small-scale and city-level studies; and how we interpret non-linearity where the predictor might be aggregated or disaggregated.

2 Background

2.1 Equation 1 in the literature

"Safety in numbers" exponents are estimated in analyses of road-traffic injuries through fitting a regression model such as Equation 1 or Equation 2 to data. At minimum, these data consist of counts of road injuries, and of two road user types, often cars and pedestrians or cyclists. The exponents β_1 and β_2 are estimated, and reported with confidence intervals and p-values corresponding to the probability of observing the data under the null hypothesis $\beta_1 = 0$, $\beta_2 = 0$.

The units of measurement are not a primary concern in studies reporting safety-in-numbers effects. On the contrary, what is identified is apparently a unit-independent, scale-invariant effect that maps cumulative road usage into injury counts. For example, Miranda-Moreno et al. (2011); Geyer et al. (2006); Garder et al. (1998); Schepers et al. (2011); Nordback et al. (2014) and Leden (2002) count vehicles (usually in terms of average daily number of vehicles per unit space, though sometimes the time unit is annual or hourly). In contrast, Prato et al. (2016) and Schepers and Heinen (2013) work in terms of km. Injuries are counted as the total in one or more years. The areas considered in the studies range in size from intersections (Nordback et al., 2014) to municipalities (Schepers and Heinen, 2013) and "local authorities" (Aldred et al., 2017).

From this we see that time and space are not explicitly included in published models. Instead, it is implicit that the numbers of injuries are linear in time and space. We return to this issue later, where we see that this specification results in a base rate (α in Equation 1) that depends on e.g. the number of years of study (Section 3.3.1).

2.2 Illustration with data for England

To give some context, we present data pertaining to road-traffic injuries recorded in England in the years 2005 to 2015. From among these data we isolate all injuries to cyclists that occurred in events involving at least one car. Alongside these data, we use Road Traffic Statistics estimates of distance travelled by cars and bikes in these areas[1], as well as census estimates for population numbers in the year 2011.

We begin with an illustration of how injuries to cyclists (involving cars) scales with the total distance travelled by cyclists (Figure 1). In this Figure, we show side by side how the scaling changes as we include only more severe cases. We repeat the analysis with London boroughs only in Figure 15.

We look also at many possible relationships between single predictors and injury counts (Figure 16). Note the similar trend as the less serious casualties are excluded (KSI: killed or seriously injured). These values are summarised in Table 2, where we include also the result of a regression of the form of Equation 1. Note that the sum aligns with the other rows.

Table 2: Coefficients

Predictor	All injuries	KSI	Fatalities
Total travel (β)	0.45	0.6	0.85
Population (β)	0.74	0.88	0.99
Bike (β)	0.68	0.77	0.86
Car (β)	0.44	0.59	0.85
Bike+Car ($\beta_1 + \beta_2$)	0.66	0.78	0.93

[1]We use data covering all road types. Meagre attempts to remove motorway travel for cars have not changed the general picture that emerges.

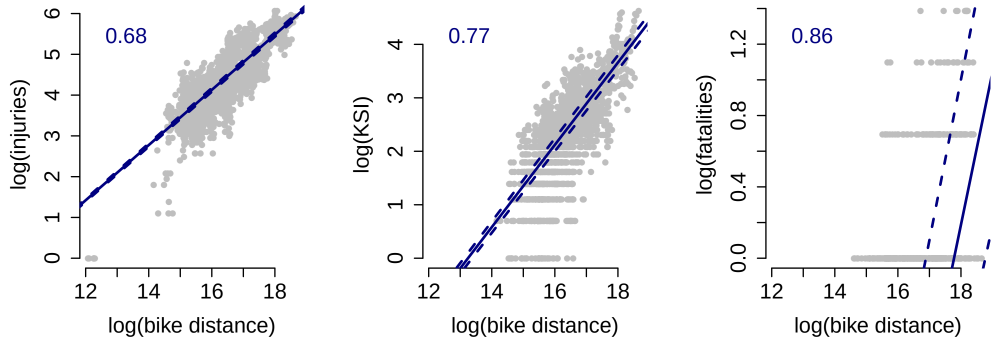

Figure 1: Fit of equation $I = \alpha C^{\beta}$ to data for local areas in England. There are 148 areas including counties and London boroughs, and 11 years (grey points). I is (left) total incidents involving a car in which a cyclist was injured; (middle) the subset of these that were KSI; and (right) the subset of these that were fatal. C counts the total distance travelled in the area. In blue is the line of best fit, with the gradient of the line (the value for β) in the top-left corner.

2.3 The scope of this work

We focus our attention on city-level models, rather than small-scale models. Our reasons for doing so are (a) brevity, (b) relevance to our aims (question 2b of Table 1), and (c) the ready availability of test data (Section 2.2). We would ultimately like to join up the work presented here with studies of small-scale areas, but more groundwork is needed first to understand how the features we explore relate to that setting, and what features we have omitted are of greater importance at a small scale.

Studies of units the size of cities are easier to conceptualise and, possibly, easier to model in an intuitive way. Relationships between size, travel, population, road length and other city-level metrics will be close to linearly related. Local features that might be significant on a smaller scale are likely averaged out at the level of a city, for example the distribution of travel over time.

Therefore, although we consider how small-scale studies might be used to make predictions for the city level (Section 4.1), we do not directly address questions (1a) and (3a) of Table 1 as further work in this area is required.

3 Examination of Equation 1

In general, it is posited that $\beta_i < 1$ for an equation of the form of Equation 1 implies safety in numbers for mode i (Elvik and Bjørnskau, 2017). However, given the construction of Equation 1, we ought to expect $\beta_i < 1$ for all i for any study counting motorists, cyclists and injuries across scales, because the mode counts C and M are functions of scale. This has implications both for our interpretation and for our application of such a model.

3.1 Interpretation of Equation 1

To illustrate, consider again the example of the English counties, excluding London boroughs, and cyclist KSI counts. Applying Equation 1, we find coefficients $\beta_1 = 0.15$ and $\beta_2 = 0.66$, and we might interpret this as safety in numbers.

However, there is an approximately linear relationship between population (N) and cyclist KSI, between population and car distance, and between population and cyclist distance (Figure 17). If we consider population to be akin to a "latent variable" that explains the relationships between the other three variables, and apply the equation $I = \alpha N^{\beta}$, we find $\beta = 0.96$ – i.e., injuries are close to linear in population.

The act of decomposing the population into two component parts splits the predictive coefficient. We would observe a three-way split were we to consider three modes. We would see the same phenomenon were we to apply this equation to, say, second-hand car sales, where I is the number of sales, and we divide the population N into M vendors and C prospective purchasers. We would see the same in HIV infection rates, if we were to count the number of new infections I and divide the population N into the susceptible group, C, and the infectious group, M (Bettencourt et al., 2007). We wouldn't describe either of these cases as examples of "safety in numbers".

3.1.1 Methodology applied to infectious disease

Let's look more closely at the infectious disease analogy, to see what are the implications of the regression method. Let's start with hypothetical, perfect "data" that illustrate the relationship $I = \alpha C^{\beta_1} M^{\beta_2}$, yielding $\beta_1 = \beta_2 = 0.5$, exemplary safety in numbers (Table 3).

Table 3: A simple example of idealised data representing safety in numbers.

Injuries I	Motorists M	Cyclists C
1	100	10
2	200	20
3	300	30
4	400	40

Now let's relabel the columns and consider instead the number of new infections in a year (I), the number of infectious people (C) and the number of susceptible people (M), Table 4.

So we see that for infectious diseases we also have safety in numbers. In fact, the rate of a disease (per capita) is often studied as a function of the total population, N, rather than the infectious and susceptible populations. We see from Table 5 that infections are linear in population ($I = \alpha' N$), even though there is safety in numbers for both infectious and susceptible parties.

For HIV infections, a relationship of $I \sim \alpha' N^{1.2}$ has been observed (Bettencourt et al., 2007). Given that $I = \alpha' N$ corresponds to $I = \alpha M^{0.5} C^{0.5}$, how would we write $I = \alpha M^{\beta_1} C^{\beta_2}$ for $I \sim \alpha' N^{1.2}$? Perhaps $I = \alpha M^{0.6} C^{0.6}$? Although this is pure speculation, it seems reasonable as a first guess. In any event, we would not expect either β_1 or β_2 to exceed 1. The resulting inference in the safety-in-numbers

Table 4: A simple example of idealised data representing safety in numbers applied to infectious disease, where we count the number of new infections per year as a function of the number of susceptible and the number of infectious people.

New infections I	Susceptible M	Infectious C
1	100	10
2	200	20
3	300	30
4	400	40

Table 5: A simple example of idealised data representing safety in numbers applied to infectious disease, where we count the number of new infections per year as a function of total population.

New infections I	Susceptible M	Infectious C	Population N
1	100	10	110
2	200	20	220
3	300	30	330
4	400	40	440

framework is then that the biggest susceptible populations M are the safest. However, the biggest susceptible populations are in the locations with the biggest populations, which have, according to the observation $I \sim \alpha' N^{1.2}$, the highest rates of new infections per year per capita.

3.2 Application of Equation 1

We ought also to consider the consequences of colinearity in using our models to make predictions. It is well known that the validity of predictive models worsens as one departs from the training space; all the more so with colinear variables, whose predictive performance is poor even within the training space (Kiers and Smilde, 2007). In addition, in many scenarios we are likely to consider in health-impact modelling, we are particularly interested in mode shifts (Schepers and Heinen, 2013): that is, we are considering transitions that break the linearity present in the construction of the model, increasing one mode and decreasing another, rather than increasing or decreasing both together (Figure 18). We return to these dynamics with a particular focus on density in a simulation study in Section 5.

3.3 Scalability of Equation 1

Equation 1 does not scale (unless $\beta_1 + \beta_2 = 1$). In part, this is due to working in numbers rather than density: the equation does not distinguish between an increase in number due to extended measurement and an increase in number due to there being more road users. The consequence is that each parametrised model is particular to its own setting, which makes the universality of finding a "safety in numbers" effect suspicious. An illustration and an example are given in Sections 3.3.1 and 3.3.2.

In fact, the absence of accounting for scale might be what gives rise to the "safety in numbers" observation in the first place. Formally, we start with a single observation $I = \alpha M^{\beta_1} C^{\beta_2}$, and make a second observation identical to the first, yielding $2I = \alpha(2M)^{\beta_1}(2C)^{\beta_2}$. We can solve these equations together to learn about the β values where we have assumed linearity:

$$\frac{2I}{I} = \frac{\alpha(2M)^{\beta_1}(2C)^{\beta_2}}{\alpha M^{\beta_1} C^{\beta_2}}; \tag{3}$$

$$2 = \frac{2^{\beta_1} M^{\beta_1} 2^{\beta_2} C^{\beta_2}}{M^{\beta_1} C^{\beta_2}}; \tag{4}$$

$$= \frac{2^{\beta_1} 2^{\beta_2}}{1}; \tag{5}$$

$$= 2^{\beta_1 + \beta_2} \tag{6}$$

So $\beta_1 + \beta_2 = 1$ when injuries are linear across a study.

3.3.1 Illustration 1

We simulate a study of observations made over different numbers of years. Starting with Equation 1, we set $\alpha = \exp(-6.879)$, $\beta_1 = 0.591$, and $\beta_2 = 0.32$, and simulate data for a single study. To simulate many studies that vary by duration, we simply multiply the data by the size of the study. We learn the parameters through a regression model corresponding to Equation 1.

Figure 2 shows that, as the size of the study increases, the base rate, α, also increases, even though the data are exactly the same. This identifies the trade-off between the parameters β_i and α: if β_1 and β_2 are fixed, our base rate α will increase as the size of the study increases. My intuition is that, in this trade-off, it is the rate α that should stay constant, while the exponents might vary with study size (Figure 3).

3.3.2 Example: England injury data

Our example uses the STATS19 injury data for England in the years 2005 to 2015, across 148 areas (which include counties and local authorities). Areas are grouped into nine regions. We apply Equation 1 (a) to area-level data within regions and make predictions for each region, and (b) to area-level data for the whole country and make predictions for the country.

Figure 4 shows the coefficients we infer for cyclist KSI counts resulting from collisions with other cyclists, motorcyclists, cars, vans, buses and heavy-goods vehicles. Empty circles are region estimates based on the areas they contain. In orange are the country estimates using all areas. In turquoise are the country estimates using the nine regions. Note the trend of clustering towards the line $\beta_1 + \beta_2 = 1$.

We now use the coefficients to make predictions at higher scales (Figure 5). There is a bias in underestimating mixed-mode injuries and overestimating same-mode injuries. In particular, we observe ~16,000 injuries to cyclists caused by cars in the country data, but predict only ~8,000.

The reason for the discrepancy is the deceptive universality of the inferred values for coefficients β_1 and β_2. While the persistence of observed values for β_1 and β_2 has been proposed as evidence that

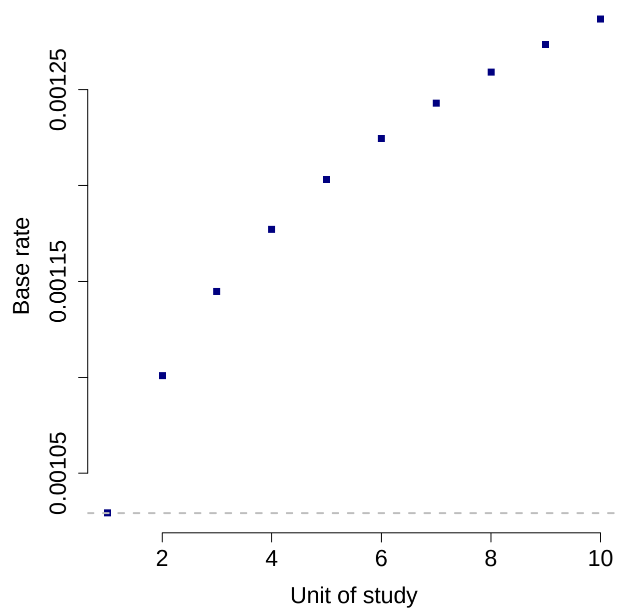

Figure 2: Base rate (α) as the size of a study increases, and safety-in-numbers exponents are fixed. The grey dashed line shows the true base rate.

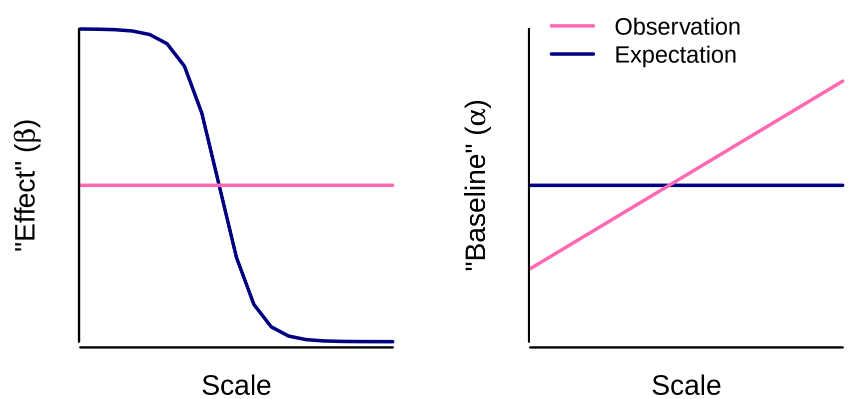

Figure 3: Safety-in-numbers effect (β) and base rate (α) as the size of a study increases. Studies observe the effect to stay more or less constant across scales, while the base rate increases (pink). One might expect instead the effect to diminish across scales and the base rate to stay constant (navy blue).

"safety in numbers" is a phenomenon, it is, to me, suggestive that it isn't. If the observed effect is independent of scale, i.e. it does not diminish as the scale increases, then we can rule out a "local protective effect" that might be conferred between proximal cyclists. If we extend the scale to include arbitrarily large areas in space and time, we can rule out policy and infrastructure effects. Then the "truth" that is conveyed by consistent discovery of values is not a feature of safety but rather a feature of the data as described by the model.

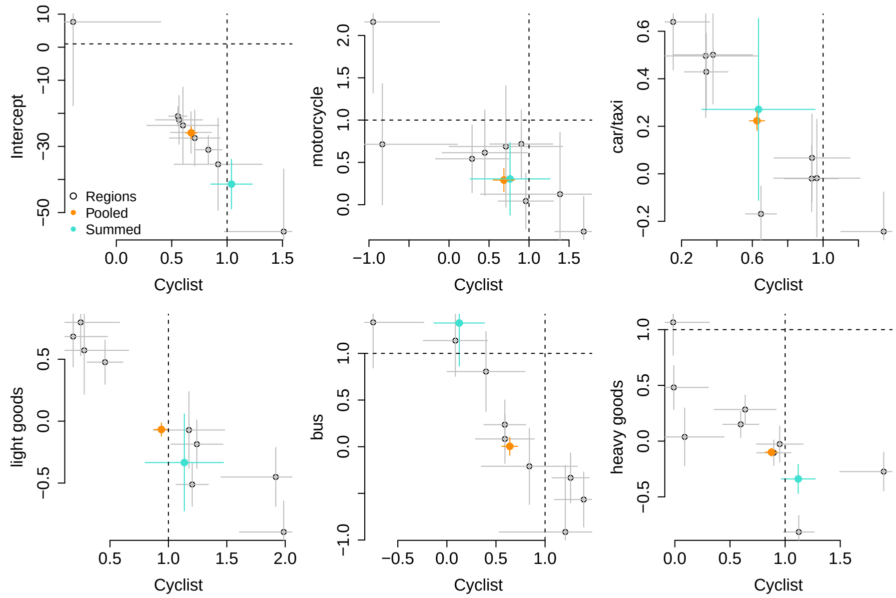

Figure 4: Coefficients we infer for cyclist KSI counts resulting from collisions with other cyclists, motorcyclists, cars, vans, buses and heavy-goods vehicles. Empty circles are region estimates based on the counties they contain. In orange are the country estimates using all counties. In turquoise are the country estimates using the nine regions.

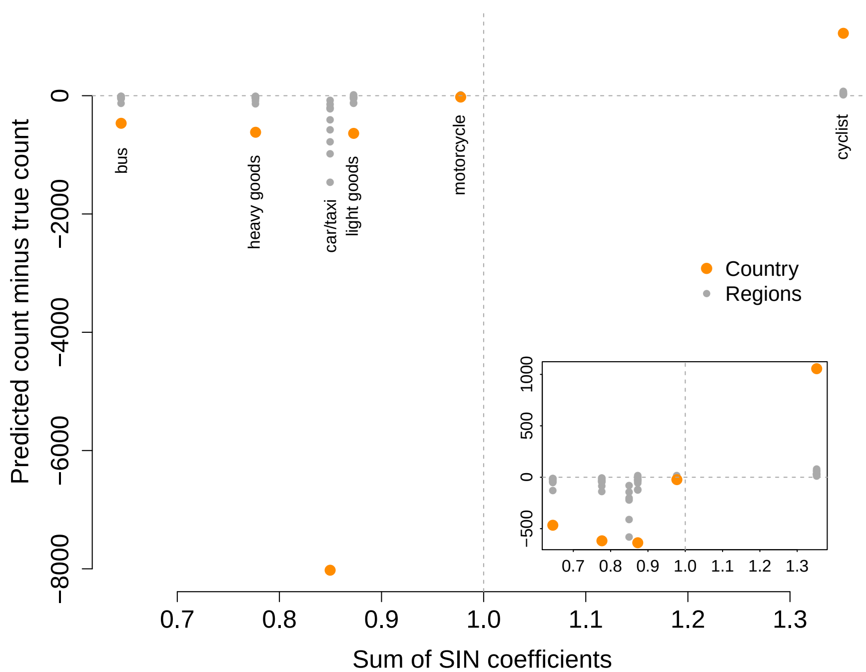

Figure 5: Predictions we make for cyclist KSI counts relative to observed counts, resulting from collisions with other cyclists, motorcyclists, cars, vans, buses and heavy-goods vehicles.

4 Application of learnt coefficients in new models

We are particularly interested in applying learning from regression modelling to making predictions. First, we explore the possibility of applying coefficients learnt at smaller scales to city-level prediction models. We describe a general method and discuss issues remaining to be resolved.

4.1 Scaling

How can we use the results of studies such as those documented in Elvik and Bjørnskau (2017) in a meaningful, appropriate way in city-level prediction models? We can define an area of effect from the studies, and extrapolate to the area for which we are predicting. This is in order to be consistent with the assumptions of the studies whose results we apply. The studies implicitly define a sphere of influence in that mode users confer protection to others within their area and not those in other areas. Therefore, in applying these models, it is consistent to employ a similar or the same area as the sphere of influence, and then scale up. Note that to use non-linear coefficients in sizes different from that of the study area defined is a violation of the assumptions inherent in the construction of the regression model.

We define the area of our unit of analysis as A, and our area of application as nA (e.g. n years). We write the total number of motorists as M and the total number of cyclists C. We approximate the numbers of motorists and cyclists in each unit A as M/n and C/n respectively. We calculate the number of injuries, I_n, as the "sum" over n identical units:

$$I_n = \alpha n \left(\frac{M}{n}\right)^{\beta_1} \left(\frac{C}{n}\right)^{\beta_2}. \tag{7}$$

Equivalently, we can calculate updated exponents as follows:

$$\beta_1' = \frac{(0.5 - \beta_1)\log(n) + \beta_1 \log(M)}{\log(M)} = \frac{(0.5 - \beta_1)\log(n)}{\log(M)} + \beta_1 \tag{8}$$

Then we would have, as in Equation 1,

$$I_n = \alpha M^{\beta_1'} C^{\beta_2'}. \tag{9}$$

This correction fixes the error of Figure 5; see Figure 6.

While this allows us to predict injuries for a whole area through consideration of small subunits, it still needs to be extended to model mode shifts and density changes in that larger region. In Section 5 we will join this up with the bigger picture.

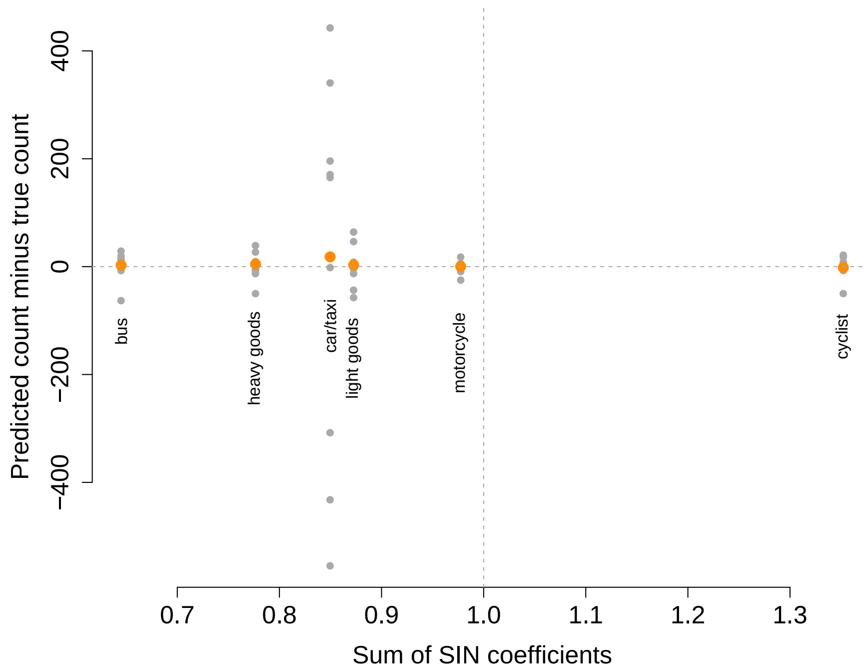

Figure 6: Predictions we make for cyclist KSI counts relative to observed counts, resulting from collisions with other cyclists, motorcyclists, cars, vans, buses and heavy-goods vehicles, as in Figure 5, but with corrected coefficients using Equation 8.

4.2 Mode definitions

This model is not robust to arbitrary redefinition of modes, e.g. we fit different models if we consider that both "red buses" and "blue buses" are the same "mode", vs. considering that they are different modes.

Let W be the number of red buses, Y the number of blue buses, and $Z = W + Y$ the total number of buses. For collisions with cycles, using equations of the form of Equation 1, we have

$$\lambda_W = \alpha_W W^{\beta_W} C^{\beta_2}, \tag{10}$$
$$\lambda_Y = \alpha_Y Y^{\beta_Y} C^{\beta_2}, \tag{11}$$
$$\lambda_Z = \alpha_Z Z^{\beta_Z} C^{\beta_2}. \tag{12}$$

Let's assume there are three times the number of red buses as blue buses, assume they have the same rate of causing injury to cyclists, and use some simple numbers to illustrate, given in Table 6.

Table 6: Idealised bus and cyclist injury data.

Observation	C	Buses (Z)	I_Z	Red (W)	I_W	Blue (Y)	I_Y
1	10	80	4	60	3	20	1
2	20	160	8	120	6	40	2
3	30	240	12	180	9	60	3
4	40	320	16	240	12	80	4

As with the illustration in Section 3.1.1, we can fit these models simply, finding in each case that $\beta_i = 0.5\,\forall i$. The intercepts, i.e. the base rates, are:

$$\alpha_W = 0.12, \tag{13}$$
$$\alpha_Y = 0.07, \tag{14}$$
$$\alpha_Z = 0.14. \tag{15}$$

It appears that both bus categories are "safer" when considered alone, with blue buses being safer than red buses, despite all bus categorisations posing the same actual danger.

We arrive at the same conclusion by defining $\exists\alpha : \alpha = \alpha_W = \alpha_Y = \alpha_Z$, reflecting our definition that all bus mode categorisations pose the same risk, and testing whether $\lambda_Z = \lambda_W + \lambda_Y$:

$$\lambda_W + \lambda_Y = \alpha_W W^{\beta_W} C^{\beta_2} + \alpha_Y Y^{\beta_Y} C^{\beta_2}, \tag{16}$$
$$= (\alpha_W W^{\beta_W} + \alpha_Y Y^{\beta_Y}) C^{\beta_2}, \tag{17}$$
$$= (\alpha_W W^{0.5} + \alpha_Y Y^{0.5}) C^{0.5}, \tag{18}$$
$$\stackrel{\exists\alpha}{=} \alpha (W^{0.5} + Y^{0.5}) C^{0.5}, \tag{19}$$
$$\neq \alpha (W + Y)^{0.5} C^{0.5} = \lambda_Z. \tag{20}$$

In order to use this implementation for health-impact modelling, it is necessary to provide a solution to this curious problem. The real analogue of this example is the combination and disaggregation of modes in injury modelling. Buses are variously combined with trucks or minibuses; cars and taxis can be one category or two; scooters, motorbikes and three-wheelers can also be considered together or independently.

4.3 How many modes?

Our predictions are highly sensitive to the number of modes we include in the model. Just as we can arbitrarily redefine buses as "red buses" and "blue buses", we can arbitrarily assume that an injury rate depends on one mode or two. This has profound consequences for our prediction methodology.

To simplify exposition, we assume that injuries modelled as a function of one mode are linear in that mode, and injuries modelled as a function f of two modes are linear in the product of those two modes. More explicitly,

$$I = f(C) = \alpha_c C, \tag{21}$$

$$I = f(M, C) = \alpha_{mc} M^{0.5} C^{0.5}, \tag{22}$$

where we use subscripts c and mc to distinguish the two models. Note that these models are both consistent with the toy example in Table 3.

The problem arises in our predictive model when we predict injuries based on one mode alone, and keep another constant. Suppose we predict the number of injuries to cyclists due to collisions with trucks. We do not change truck distance in our scenario, so a model of the type of Equation 21 seems appropriate, as we are predicting the number of cyclist injuries as a function of cyclists, not of cyclists and trucks. However, they produce very different predictions for the case of doubling cyclists: $2I$ injuries predicted from Equation 21, vs. $\sqrt{2}I$ injuries predicted from Equation 22.

What are the consequences of this insight for predictive models where one mode changes not at all, or only a little? Does this bring us back to the problem discussed in Section 3.3? If yes, how ought we rewrite e.g. Equation 8 to account for the varying dependence of the function on the mode arguments? We explore some of these questions in the next section (Section 5).

5 Simulation and motivation for a new model

Here we look explicitly at city-level models, that is, studies that consider multiple cities (rather than small localities such as junctions) in order to understand how safety scales as a function of road use. As at other scales, coefficients $\beta_1, \beta_2 < 1$ are understood to correspond to safety in numbers and, again, coefficients learnt in such studies are suggested as parameters to be used in predictions of injury numbers in mode-shift scenarios.

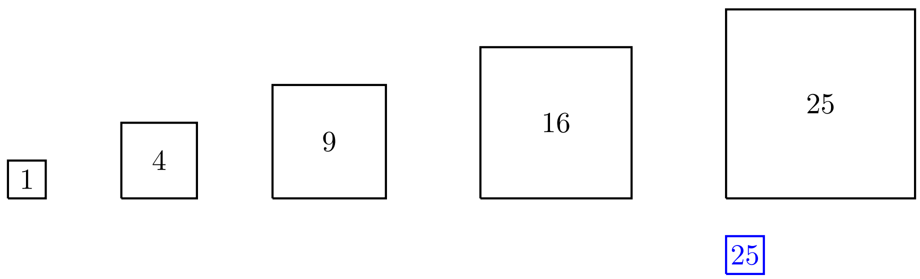

Figure 7: Schematic representation of an inter-city study, demonstrating scaling across cities. If we fit a model using these cities, and then make a prediction for "City 1" with 25 times the travel, then the best estimate we make will correspond to the observation for "City 25" (black). However, when modelling mode-shift scenarios, what we actually want to predict is represented in blue: the total travel of "City 25", taking place in a city of the size of "City 1".

Figure 7 demonstrates the fundamental problem in using cities that vary across scales to parametrise models intended for use in predicting for a city that will change in number but not scale, i.e. it will change in density. In this section we develop the ideas summarised in Figure 7 through a simulation study.

5.1 Simulation model

We create a simulator with three variables: the number of cyclists, the number of motorists, and the dimension of the space, which we equate to the size of the city. Each cyclist and each motorist is a five-pixel by five-pixel square. At each time step each body moves one body-width. We simulate 500 time steps. Each body undertakes a biased random walk: that is, with probability 5/6 it continues in the same direction, and with probability 1/6 it chooses a direction randomly from among the directions available to it (including the same direction). Moves that would take a body out of the frame wrap around to the other side of the frame. The frame size is some multiple of the step size that we vary as an input parameter.

We count the number of collisions between cyclists and motorists, which is defined as any overlap in pixels between a cyclist and a motorist. In the style of PacMan, upon collision the cyclist disappears and reappears randomly in the next time step. The cyclist disappears immediately, so a collision involving two motorists and one cyclist will be counted as one event.

5.2 Simulated study

In order to emulate inter-city regression studies, we simulate 50 times frames of different sizes and constant density. We define our vector of sizes to be the integers $x = \{5, ..., 14\}$. Then the numbers of cyclists and motorists are Poisson-distributed random variables with mean x^2, and the dimensions of the frames are $20 \times 5x$ pixels by $20 \times 5x$ pixels. There are 500 simulations in total: ten sizes with fifty repetitions of each. The results of the simulations are summarised in Figure 8. We fit a Poisson regression model to the resulting number of collisions to learn the parameters β_1 and β_2, finding

$\beta_1 \approx \beta_2 \approx 0.5$ as expected. We then use (a) the model to make new predictions and (b) the simulator to test them.

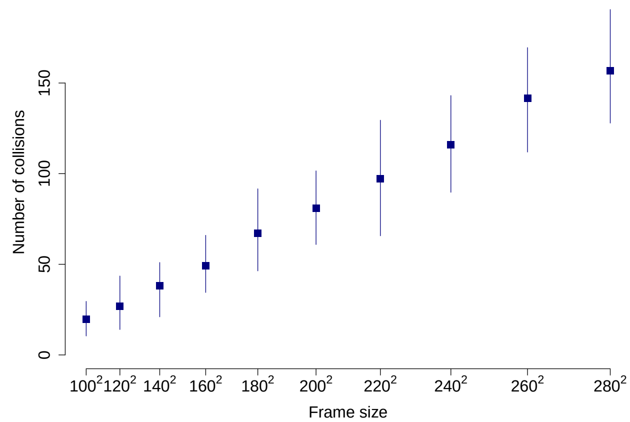

Figure 8: Simulation study. The number of collisions is a function of frame size, number of cyclists and number of motorists. Each bar shows 90% of the range of 50 simulations. In every simulation the density of cyclists and the density of motorists is the same.

5.3 Comparison of model predictions to simulations

First, we consider every frame size with 100 motorists and 100 cyclists. This was one of our inputs into the regression model data: frame size 200^2. Because our model is a function of road-user numbers alone, we make the same prediction for every frame size: 80 collisions. In Figure 9 we plot this against the simulated number over 50 repetitions for each frame size. Note that only for the frame size corresponding to mode numbers of 100 does the 90% confidence range overlap the predicted value.

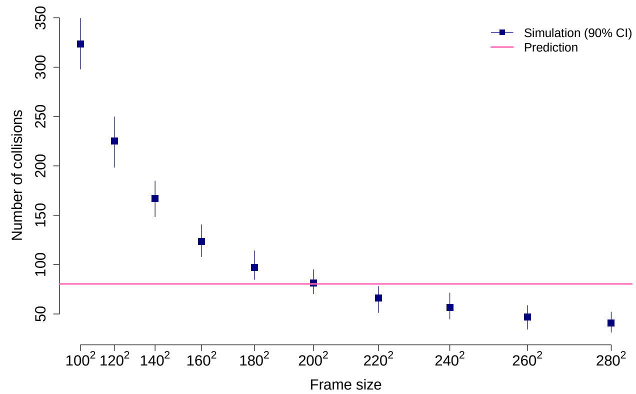

Figure 9: Simulated vs. predicted results for 100 motorists and 100 cyclists with varying frame sizes. The predicted value is 80 for all frame sizes.

Second, using the same set up, we consider that the number of motorists is constant at 100 in the corresponding frame size of 200^2, and we vary the number of cyclists. Again, we predict the number of collisions using our model, which this time varies with cyclist number. Note in Figure 10 that the prediction aligns with the simulation when the cyclist value corresponds to the frame size and density of the predictive model. When this number is exceeded, the estimate is too low, and at a lower density, the prediction is too high. We should keep this picture in mind when we consider e.g. cyclist injuries when cyclist numbers are changing but the other mode, such as truck, is not.

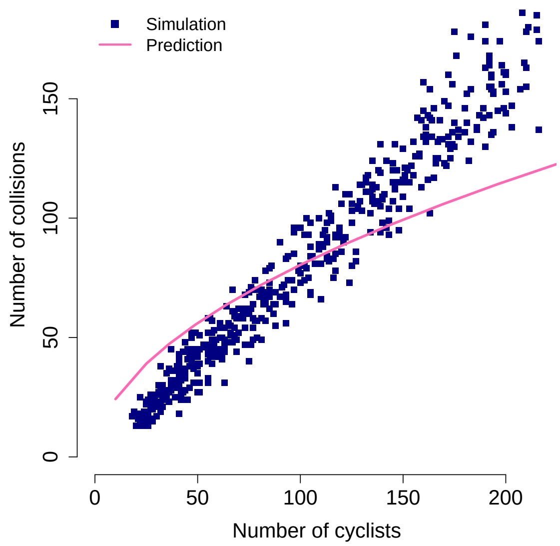

Figure 10: We fix motorists at 100 and frame size at 200^2, and vary the number of cyclists as Poisson random variables with means from 25 to 196. We simulate and predict using the model the number of collisions. This represents the type of error we might make if we predict for one mode that varies when another stays constant.

Third, we consider that the number of cyclists is constant at 150 in a frame size of 200^2, and we vary the number of motorists. Again, we predict the number of collisions using our model, which this time varies with motorist number. Note in Figure 11 that the prediction aligns with the simulation when the motorist value corresponds to the difference between the motorist value corresponding to the frame size and density of the predictive model, and the change in cyclist number from its corresponding value. (Precisely: the errors cancel out when the number of new motorists = the number of old motorists $\times$ the number of old cyclists / the number of new cyclists $= 100 \times 100/150 = 66.7$.) When this number is exceeded, the estimate is too low, and at a lower density, the prediction is too high. We should keep this picture in mind when we consider e.g. that we have six modes each with 100 road users, and we predict for a scenario in which cyclists increase to 150 and all other modes decrease to 90.

5.4 Simulated safety in numbers

We repeat the simulation study, this time imposing a safety-in-numbers effect. We specify that the probability of collision is represented by some number $p = p(C)$, which is a function of the number of cyclists C. We recreate Figure 8 with $p = C^{-0.25}$. In terms of "safety in numbers", this corresponds to a raw exponent of 0.75. The probability p doesn't depend on M, so its exponent is 1.

We see the effect of cyclist number on collision number in Figure 12. Note the non-linear gradient, compared to Figure 8. For these data, we infer for a model of the form of Equation 1 scaling exponents $\beta_1 + \beta_2 \approx 0.75$, which is equal to the sum of the raw exponents minus 1.

Next, we predict the number of collisions for this model with doubled numbers of cyclists and motorists and varying frame sizes. The results are shown in Figure 13 along with the prediction from the the scaling exponents inferred from Figure 12.

Concatenating data in Figures 12 and 13, we fit a model with scaling exponents $\beta_1 + \beta_2 \approx 1$, recovering the tiling parameters. When we control for frame size (i.e. we specify a different model, which eliminates confounding), we isolate the density effect: we will refer to these exponents (those

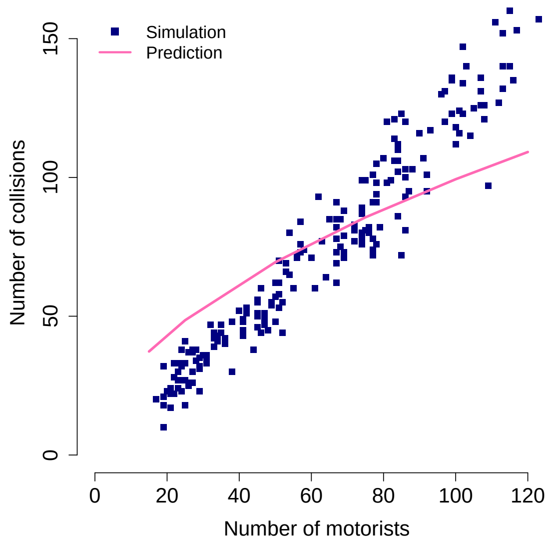

Figure 11: We fix cyclists at 150 and frame size at 200^2, and vary the number of motorists as Poisson random variables with means from 25 to 100. We simulate and predict using the model the number of collisions. This represents the type of error we might make if we predict for one mode that varies as a result of reallocation of other mode types. Note the deviation between model and prediction, which is minimised close to the point of constant density, i.e. that the increase in cyclists is close to the decrease in motorists.

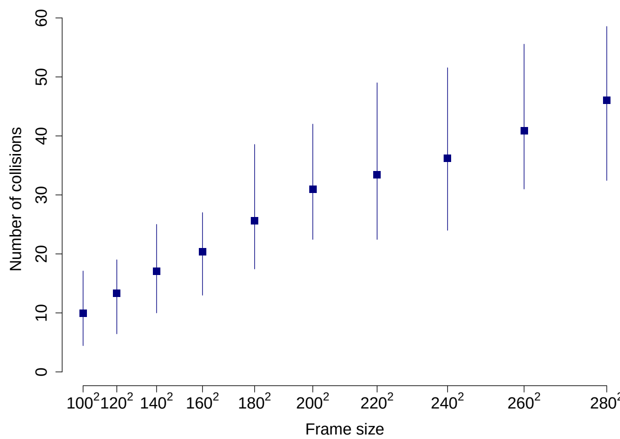

Figure 12: Simulation study. The number of collisions is a function of frame size, number of cyclists and number of motorists. Each bar shows 90% of the range of 50 simulations. In every simulation the density of cyclists and the density of motorists is the same. Collisions are a function of the number of cyclists and occur with probability $p = C^{-0.25}$.

belonging to the density model) as δ_1 and δ_2, analogous to β_1 and β_2. We find that $\delta_1 + \delta_2 \approx 1.78$, close to the original exponents.

We repeat the whole process for $p = C^{-0.5}$, i.e. exponents 0.5 and 1. Analogously for Figure 12 we find $\beta_1 + \beta_2 \approx 0.5$, the sum of the original exponents minus one. Again we simulate data for doubled cyclists and motorists which, concatenated to the original data, yield scaling exponents $\beta_1 + \beta_2 \approx 1$ and density exponents $\delta_1 + \delta_2 \approx 1.5$ when size is accounted for.

We are trying to build up a picture of how Equation 1 behaves across scales and densities. It seems that the predictive equation we are aiming for requires density exponents; that studies likely return

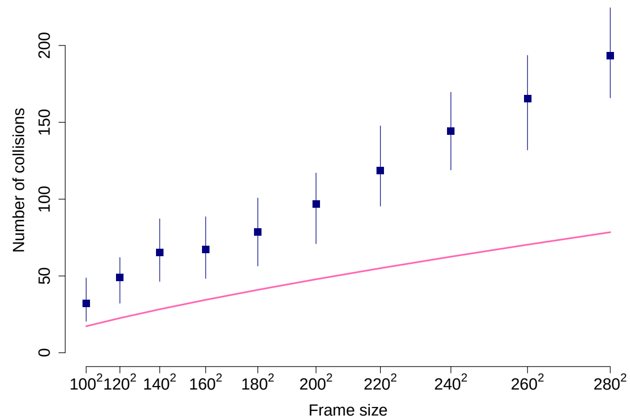

Figure 13: As in Figure 12, with double the number of motorists and cyclists. The prediction generated from data in Figure 12 is shown in pink.

scaling exponents; and that, for two modes, the sum of the density exponents is one more than the sum of the scaling exponents: $\delta_1 + \delta_2 \approx \beta_1 + \beta_2 + 1$.

5.5 Discussion

We do not simulate in order to suggest that we have understood and can recapitulate the mechanism. Rather, we use a simple simulation model in order to test simple assumptions and infer simple principles, for example the relationship between size and density. If we can make a general statement about what happens in this simple simulation, in which we have full control of the workings, we have the opportunity to illuminate possible factors influencing observational inferences such as those in safety-in-numbers studies.

There are some key differences between our simulation study and real data: we have no spatial or temporal factors, and no relationship between e.g. density and speed. One respect in which our simulation differs from the England study of Section 3.3.2 is that we consider here constant density across scales, whereas, on average, the areas in England decline in density as size increases. In contrast, in general, it is posited that, globally, as city size increases, so does density (see schematic of the model space, Figure 14). However, we could tailor our simulation to match in some way a phenomenon we are interested to capture.[2]

The point here is not how much our simulation set up resembles what one imagines happens in cities. The point is that we can contrive data from which we can learn coefficients that fit the general trend $\beta_1 \approx \beta_2 \approx 0.5$, and we can test this model against data simulated from the same original source. This means we can test the range of applicability of the model.

5.6 Formal presentation of an alternative model

We can formally summarise the model used in place of Equation 1 to correctly capture the dynamics in the simulations for general use. For n units, with $c = C/n$ cyclists per unit and $m = M/n$ motorists per unit, we write:

$$I_n = \frac{\alpha}{n}(nm)^{\delta_1}(nc)^{\delta_2} = \frac{\alpha}{n}M^{\delta_1}C^{\delta_2} \tag{23}$$

[2] For example, we could introduce density-dependent speeds. Concretely, in a fixed space, doubling cyclists and doubling motorists increases density by a factor of four. If we then divide every mode speed by that value - four - we undo exactly the effect of the increased density.

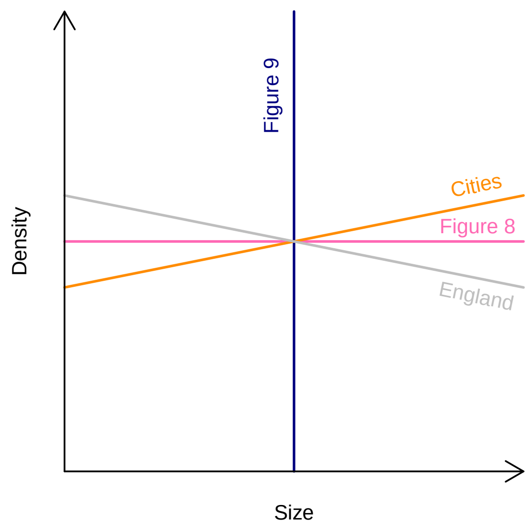

Figure 14: A depiction of the model space covered by the simulation model. We mark out in pink and navy blue a line of constant density and a line of constant size through the model space, depicted in Figures 8 and 9, respectively. The study of areas in England might be represented in this model as occupying the space of increasing size and decreasing density (grey), and expect that a study of many cities will occupy the space of increasing size and increasing density (orange). (NB: these are schematic.)

where n is the number of spatial units. Then, in a scenario, we can distinguish between a doubling of density,

$$I_n^{(\text{double density})} = \frac{\alpha}{n}(2M)^{\delta_1}(2C)^{\delta_2} \tag{24}$$
$$= 2^{\delta_1+\delta_2}\frac{\alpha}{n}M^{\delta_1}C^{\delta_2} \tag{25}$$
$$= 2^{\delta_1+\delta_2}I_n, \tag{26}$$

and a doubling of size:

$$I_n^{(\text{double size})} = \frac{\alpha}{2n}(2M)^{\delta_1}(2C)^{\delta_2} \tag{27}$$
$$= 2^{\delta_1+\delta_2-1}\frac{\alpha}{n}M^{\delta_1}C^{\delta_2} \tag{28}$$
$$= 2^{\delta_1+\delta_2-1}I_n. \tag{29}$$

With this formulation we can also derive the relationship with β_1 and β_2, assuming observations for a size of 1 and a size of n. From Equation 23 we have

$$\frac{I_n}{I_1} = \frac{\frac{\alpha}{n}(nm)^{\delta_1}(nc)^{\delta_2}}{\alpha m^{\delta_1}c^{\delta_2}} \tag{30}$$
$$= n^{\delta_1+\delta_2-1}. \tag{31}$$

We can derive the equivalent relation using Equation 1:

$$\frac{I_n}{I_1} = \frac{\alpha(nm)^{\beta_1}(nc)^{\beta_2}}{\alpha m^{\beta_1}c^{\beta_2}} \tag{32}$$
$$= n^{\beta_1+\beta_2}. \tag{33}$$

Thus we see that for these two models to be consistent, we require $\beta_1 + \beta_2 + 1 = \delta_1 + \delta_2$.

5.7 Conclusion

The conclusion of this test is that a predictive model that does not take into account factors of scale fails to predict outside its training space. Further, we can identify the direction of the bias: uncaptured increases in density lead to underprediction of collisions, and uncaptured decreases in density lead to overprediction. This calls into question the assumption that density-independent regression models can be used to predict the number of collisions, or injuries, that will occur in mode-shift scenarios.

A heuristic for correction is offered: for our simulation, scaling exponents of 0.5 corresponded to density exponents of 1, with two modes simulated. A starting point for application of inter-city studies is then the translation of a scaling exponent of β_i to a density exponent of $\delta_i = \beta_i + 0.5$, where the number of modes we model is two. This would correct the errors seen in Figures 9, 10 and 11.

Recall that in Section 4.1, we found that, for two modes (and an estimated n units of space) $\lim_{n\to\infty} \beta_i = 0.5$, which corresponds to the size scaling we have just explored. Then our full pipeline, from small-scale study to exponents for injuries as a function of distance, where density is taken into account, would be:

1. Identify small-scale exponents
2. Calculate city-level scaling exponents using Equation 8
3. Use city-level scaling exponents to calculate city-level density exponents by adding a total of 1 to the parameters, e.g. 0.5 to each.

Then our null hypotheses are, for models of the form of Equation 1, that $\beta_1 = \beta_2 = 0.5$ for constant density and varying scale, and $\beta_1 = \beta_2 = 1$ for constant scale and varying density. These models can be captured together in a way that is mutually consistent in a model parametrised by exponents δ_1 and δ_2 and also by size; i.e. they are "special cases" and complementary presentations of the overarching model.

We propose the model of Equation 23 as a model that is consistent with our observations and our insights. There will be other models that also fulfill those criteria. There will be other models that contain ours within them as a subset or a special case. An example would be an extension that takes account also of speed. Whatever model we use, however, ought to be consistent with the observations made so far, and this is one such model.

In terms of application, we don't know how to add 1 to β_1 and β_2. We could simply define $\delta_i = \beta_i + 0.5$, adding an equal amount to each variable. However, inference of β_1 and β_2 is likely confounded by space and/or time, so we might instead choose $\delta_1 = \delta_2 = (\beta_1 + \beta_2 + 1)/2$. This is an equitable solution. These should be tested, particularly for cases where one mode vastly outnumbers the other. For models expressing uncertainty, distributions can be assigned to these parameters and a sensitivity analysis conducted to ascertain the impact of the values on the outcome.

Finally, the problem identified in Section 4.3 remains for $\delta_1 \neq 1$ where δ_1 is applied to a mode that is arbitrarily redefined. A quick fix would be to define $\delta_1 = 1$ for this mode, or group of modes, and $\delta_2 = \beta_1 + \beta_2$. Of course this won't help when both modes are arbitrarily defined. In this case, one can re-parametrise the model, or make multiple assessments (one per mode definition) as a sensitivity analysis.

6 A density study of the England data

We apply the previously presented model (Equation 23) to the England data, using the distance travelled and the total road length for each area, assuming a Poisson distribution for the counts. We use the 148 areas of England and consider only A, B and minor roads. We consider urban and rural areas both separately and together, defining urban areas as those with at least 98% of their population registered as living in a city, town or minor conurbation in the 2011 census (2079 data points; 2889 for rural). We use the software Stan in R to test the hypothesis $\delta_1 + \delta_2 = 2$, using default (uniform) priors.

6.1 Results

In Table 7 we present the 95% credible intervals. Given our data and model, we can reject the null hypothesis of linearity ($\delta_1 + \delta_2 = 2$) for fatalities in urban areas, but not in rural areas. We reject the null hypothesis for both types of area for KSI and all injuries. The ranges are much larger for fatalities, presumably because the signal is weaker as there are fewer events and more zeros. It might be that with more data we would reject the null hypothesis for rural fatalities also.

We note that the sums are about 1 more than the sums $\beta_1 + \beta_2$ that we learn with the same data and a simpler model that considers numbers alone and not road length (Table 2 for "All areas": 0.66 for all injuries, 0.78 for KSI, and 0.93 for fatalities). (In fact, the difference is slightly greater than 1, as the areas in England have a slightly negative correlation between size and cumulative cycling, and hence greater adjustment is required, depicted as the gap between the grey and navy lines in the schematic in Figure 14, which exceeds 90°.)

Table 7: 95% credible intervals for $\delta_1 + \delta_2$.

	All injuries	KSI	Fatalities
All areas	1.73–1.75	1.84–1.87	1.88–2.01
Urban areas	1.54–1.58	1.66–1.74	1.33–1.80
Rural areas	1.88–1.89	1.93–1.97	1.86–2.03

In Tables 8 and 9 we present the 95% credible intervals for the coefficients for cycle and car travel, and in Table 10 the intervals for the intercepts. Note the negative correlations between coefficients: for "All areas" in Tables 8 and 9, as casualty severity increases, more of the coefficients' sum is attributed to car at the expense of cycle. Similarly, across all urbanicity levels, there is a negative correlation between the sums in Table 7 and Table 10, suggesting a tradeoff between base rate and safety effect.

Table 8: 95% credible intervals for δ_2 (cycle travel).

	All injuries	KSI	Fatalities
All areas	0.80–0.82	0.76–0.80	0.58–0.79
Urban areas	0.81–0.84	0.81–0.87	0.59–0.93
Rural areas	0.52–0.55	0.51–0.57	0.33–0.63

Finally, we present the coefficients we fit when we combine all datasets using factors (Table 11). Using factors for casualty severity and urbanisation, we end up with coefficients that resemble those fit to the dataset with "slight" injuries and pooled urbanisation.

Table 9: 95% credible intervals for δ_1 (car travel).

	All injuries	KSI	Fatalities
All areas	0.92–0.94	1.06–1.09	1.17–1.36
Urban areas	0.72–0.75	0.84–0.90	0.63–0.99
Rural areas	1.34–1.37	1.37–1.45	1.31–1.62

Table 10: 95% credible intervals for α (the intercept).

	All injuries	KSI	Fatalities
All areas	-21.1–-20.9	-25.7–-25.2	-32.3–-29.6
Urban areas	-17.6–-17.2	-22.8–-21.5	-27.2–-19.2
Rural areas	-25.5–-25.2	-28.9–-28.1	-34.0–-29.8

6.2 Discussion

We present the test of $\delta_1 + \delta_2 = 2$ as the test for "size-adjusted safety in numbers", which can be done in the Bayesian framework with Stan. Using glm in R, we can test $\delta_1 = 1$ and $\delta_2 = 1$ (i.e., test per capita as in Shalizi (2011)) using the offset function, but we cannot test their sum.

Both in the glm framework and with Stan, it is unclear to what extent we can interpret the values for δ_1 and δ_2. They are likely not identifiable. Is it the case that there is more safety in numbers for cyclists in rural areas, and a corresponding danger in numbers from cars in rural areas? Or is there sharing of the coefficients, so that both values should be closer to 0.95?

There appear to be correlations between coefficients across the models. What does it mean that, for "All areas" in Tables 8 and 9, as casualty severity increases, more of the coefficients' sum is attributed to car at the expense of cycle? It seems likely that there is a pattern underlying the data, which is consistent across severity levels, and that the model is unable to describe. It seems less likely that casualty severity impacts on the dynamics of the non-linear relationship between distance travelled and road-traffic collision rates. There is also a negative correlation between base rate and safety effect in Tables 10 and 7: do places that are more safe have less of a safety effect? Or is there some transference between the parameters?

This analysis is comparable to the population-adjusted model of Aldred et al. (2017). In that study, there are three covariates: cycle commuters, motor vehicle volume, and population. Here, we have two covariates, cycle distance and motor distance, and one offset, the total length of A, B and minor roads.

The number of cycling commuters is similar to estimated cycle distance; motor vehicle volume and motor distance are measuring the same thing; and population and road length are correlated, and are both a proxy for an area's size. The key difference between the two models is the treatment of the city proxy as a covariate (population) and as an offset (road length), which is equivalent to a covariate with a fixed coefficient, which we set to -1.

The sums of the means for the covariate coefficients of the population-adjusted model of Aldred et al. (2017) are 0.98, 1.06, 1.02, and 0.99. We cannot test a hypothesis in this regression framework but it's worth noting the proximity to our corresponding estimates, $\delta_1 + \delta_2 - 1$, which are consistently just less than 1.

That there is a systematic difference between "urban" and "rural" areas in terms of the coefficients fit by our model is indicative that the model does not capture the whole effect of travel density on injury rates. This, and the spurious correlations in coefficients, show there are features of the data not explained by the model of Equation 23.

Table 11: 95% credible intervals all coefficients using a factor covariate for severity.

	δ_2 (cycle travel)	δ_1 (car travel)	$\delta_1 + \delta_2$
All areas	0.80–0.82	0.92–0.94	1.74–1.75
Urban areas	0.73–0.75	0.81–0.84	1.55–1.58
Rural areas	0.49–0.54	1.34–1.39	1.88–1.89
With factor	0.79–0.80	0.97–0.99	1.77–1.78

Mis-specification of any of (a) the relationship in the model between size and rate, (b) the probabilistic description of the error term, (c) the component contributions of the two mode distances, and (d) quantification of size through road length might contribute to the failure to explain the data, and all would benefit from further consideration and testing. All of these things can be improved upon, incrementally, as we have taken an incremental step from Equation 1 to a size-based model in Equation 23. However, taken together, they highlight that it has not been shown that a model with the fundamental form of Equation 1 might be capable of answering the question of safety linearity.

7 Conclusions

For studies that scale across sizes, $\beta_1 + \beta_2 = 1$ for Equation 1 represents linearity in numbers. That the same (or similar) coefficients are observed across scales is perhaps the biggest indication that whatever is being captured is not an effect of cyclists conferring protection to other nearby cyclists.

The missing component in Equation 1 is size. Omission of size results in misleading interpretations and predictive models unsuitable for fixed settings. Therefore, we recommend departing from this model and developing new expressions, which include size explicitly. We propose, in the first instance, a very simple adjustment to Equation 1 in Equation 23, in which we include size as defined by total road length.

We recognise the challenge of specifying testable hypotheses in this setting. Ideally we would test per capita rates for δ_1 and δ_2 in Equation 23, but, as the model stands, they cannot be confidently identified, so we test instead the null hypothesis $\delta_1 + \delta_2 = 2$ in Stan. With this model, hypothesis, and the data for England, we found a size-adjusted safety-in-numbers effect for all subsets of the dataset, with the exception of the set of fatalities in rural areas, whose null hypothesis we do not reject given the data we have.

The model we present is best described as "size-adjusted safety in numbers"; we have not developed or tested a model of "safety in density", which might be interesting and relevant to explore. It would allow for heterogeneity in density over space and time and exploration of their impacts on results. That will be particularly important for small-scale studies, e.g. of junctions, and could be addressed in the first instance through simulations such as those presented in Section 5. In addition to heterogeneity in density over time and space, an improvement to the model in Equation 23 would be inclusion of mode speeds.

In terms of hypothesis testing, we aim to engage a wider audience and share data in order to find alternative ways to understand and formulate the problem. We recognise a correspondence between our model and city-level metrics that were once claimed to exhibit power-law scaling properties (Leitão et al., 2016) (see Appendix C). Input from these statistical modellers might greatly benefit progress in this topic. In addition, there might be parallels with the practice of discretisation of space (and time) in a "contact matrix" to describe interactions across partitions in infectious-disease modelling (Birrell et al., 2011). We have identified some areas for further testing or development of the model described by Equation 23, such as specification of the error term, use of density proper rather than size adjustment, the expression of the mode distances, how to quantify size, and the problem of mode (dis)aggregation. We would welcome development of other models, as well as methods for assessment and hypothesis testing.

Complementing data-driven analyses, simulation models can be used to develop relational models. These can test implications of comprehensive mechanistic models of injuries as a function of space and its occupancy. Simulation and theory provided us with a number of null hypotheses, which (a) give us an objective to test and reject, and (b) provide a justified basis for prediction. In this work, the theory and simulation led to the following hypotheses, which connect Equation 1 to Equation 23:

(1) We can estimate the city-scaling exponent from small-scale exponents by scaling β coefficients in such a way that we approximate the process of "multiplying then summing" rather than "summing then multiplying" (Section 4.1).

(2) City-level size-scaling exponents $\beta_1 = \beta_2 = 0.5$ correspond to linearity (Section 3.3).

(3) City-level size-scaling exponents $\beta_1 = \beta_2 = 0.5$ (Equation 1) correspond to city-level density-scaling exponents $\delta_1 = \delta_2 = 1$ (Equation 23), where mode speeds are assumed independent of density.

(4) To predict the consequence of a change in density using city-level size-scaling exponent β, we can use density exponents $\delta_1 + \delta_2 = 1 + \beta_1 + \beta_2$, where there are two modes involved (Section 5.4).

These hypotheses were consistent with the results of the England data, in that the coefficients δ_1 and δ_2 fit with the model of Equation 23 had sum approximately one greater than that of β_1 and β_2 fit with the model of Equation 1. We therefore propose this framework as a first amendment for making predictions of road-injury burden in mode-shift scenarios (question (2b) in Table 1). With reference to all those initial objectives, we hope to have offered a new perspective, and to have generated more questions and pointed to new lines of inquiry. We hope that new ideas and, eventually, answers will be forthcoming.

A Supplementary figures

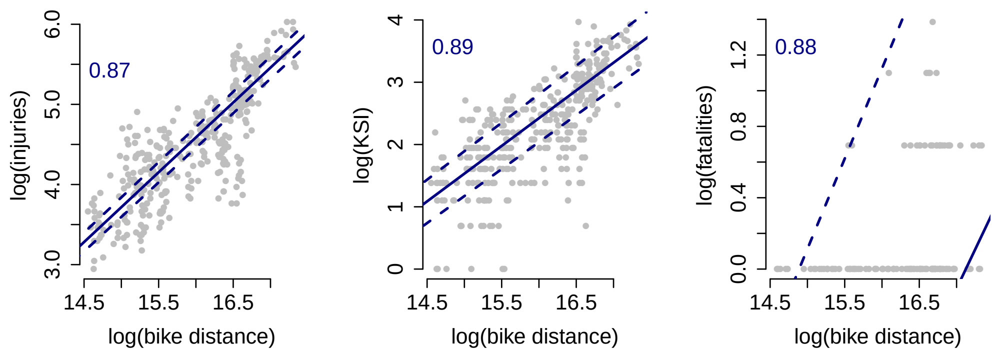

Figure 15: As in Figure 1, with only the London boroughs.

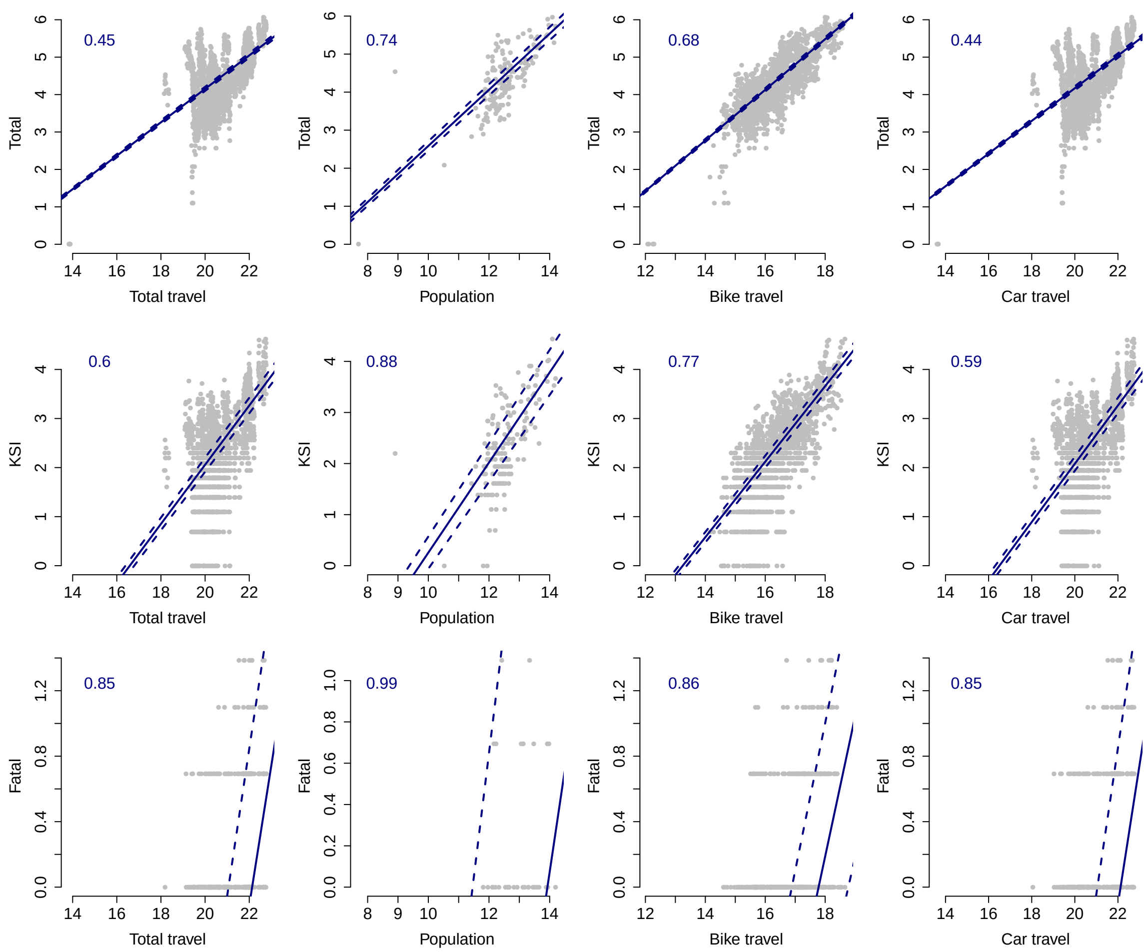

Figure 16: As in Figure 1, showing different relationships and the β coefficients they generate.

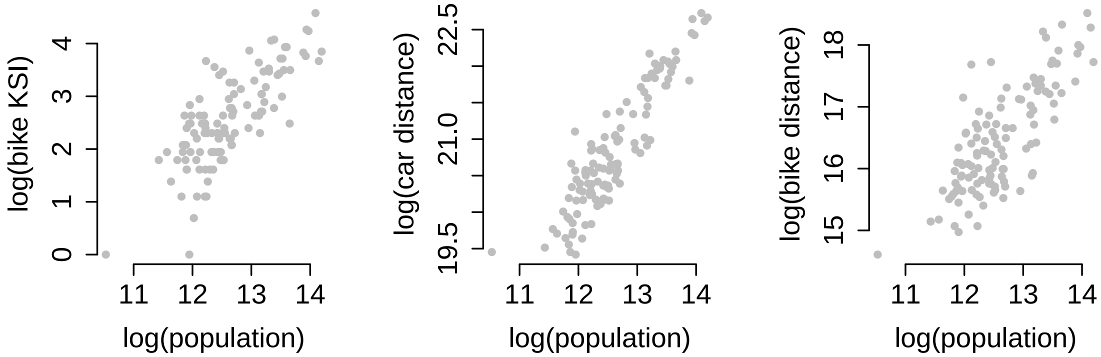

Figure 17: Linear relationships between population (N) and cyclist KSI, between population and car distance, and between population and cyclist distance in English counties.

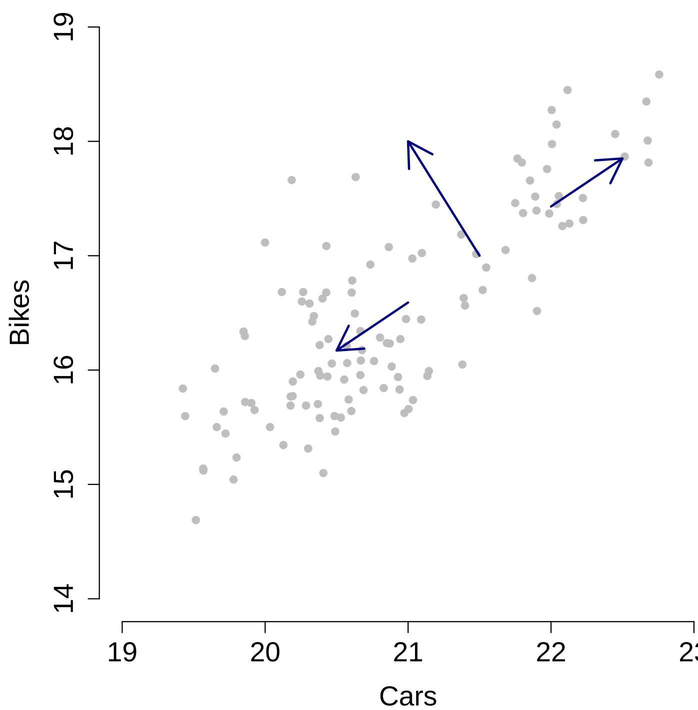

Figure 18: Where we have two colinear predictors, it seems reasonable to make predictions in spaces that align with their colinearity. However, does it make sense to make predictions in spaces orthogonal to the colinearity?

B Worked example: implementation in ITHIM-R

We consider the setting of Accra, for which we have a list of recorded fatalities over multiple years. Each record contains the following information: the year, the mode of the casualty, the mode of the other party, the age of the casualty, and the gender of the casualty. In addition, we have a travel survey, from which we learn total travel by each mode (and by demographic group, which we omit for now, for simplicity).

B.1 Constructing the model

We fit the observed data (the number of injuries, I) to an equation of the form

$$I \sim \text{Poisson}(\lambda), \tag{34}$$

$$\lambda = \alpha M^{\beta_1} C^{\beta_2} \exp\left(\sum_{i=3}^{P} X_i \beta_i\right) \tag{35}$$

with α a fixed intercept, C and M the distances travelled by cyclists and cars, respectively, based on the travel survey, and X the model matrix built from all the covariates (here, we consider only the two modes; gender and age of the casualty are omitted for simplicity). We do not use the "year" covariate but instead suppose that we have multiple observations for a single "year" (i.e. we reuse the distance data). Finally, the coefficients to fit using `glm` are α and β_i for $i \geq 3$, and we supply β_1 and β_2 as fixed parameters so that $M^{\beta_1} C^{\beta_2}$ is our offset.

Note that there are many combinations of modes, so this model is linked via the model matrix X to the number of pedestrian casualties in collisions with buses, etc. The contingency table of injury counts between all mode pairings forms the "who hit whom" matrix for the city.

B.2 Making predictions

We use the same model equation to make predictions in hypothesised scenarios. The prediction equation requires us to specify the distances travelled in the scenario: call them $\hat{M}$ and $\hat{C}$. Then we predict the expected number of injuries in the scenario, $\hat{I}$, as:

$$\hat{I} = \alpha \hat{M}^{\beta_1} \hat{C}^{\beta_2} \exp\left(\sum_{i=3}^{P} X_i \beta_i\right). \tag{36}$$

To aid interpretation, we can consider the ratio of the expected injuries in the scenario to the expected injuries in the baseline:

$$\frac{\hat{I}}{\mathbb{E}(I)} = \frac{\hat{M}^{\beta_1} \hat{C}^{\beta_2}}{M^{\beta_1} C^{\beta_2}} \tag{37}$$

$$= \left(\frac{\hat{M}}{M}\right)^{\beta_1} \left(\frac{\hat{C}}{C}\right)^{\beta_2}. \tag{38}$$

Then we can immediately read out, for example, that if M does not change ($\hat{M} = M$) then the fold change in injuries is equal to the fold change in cycling raised to the power β_2: if $\beta_2 = 1$, then if cycling increases 25 times, so does the injury count. If $\beta_2 = 0.5$, then if cycling increases 25 times, the injury count increases five times.

B.3 β_1 and β_2 parameters

The question we need to answer is, given that we are using this model, what values should we choose for β_1 and β_2? This choice will impact on the other parameters to fit (α and β_i for $i \geq 3$) and, crucially, on the number of injuries we predict in scenarios.

Recall that there are multiple casualty modes and multiple "other party" modes, including NOV (no other vehicle). Another question we need to answer is how these values should differ for different modes, in particular (a) where a mode's distance is not changing at all (or even very little) in scenarios, (b) where a mode is a combination of multiple modes, and (c) where there is no other mode.

C Power-law scaling relationships

The relationship described by Equation 1 closely resembles the family of power-law scaling models in e.g. Bettencourt et al. (2007). We could introduce this field of study to that family, describing safety in numbers as another relationship showing a power-law scaling relationship, meriting further investigation, for which solutions are occasionally proposed (Sim et al., 2015; Shalizi, 2011). This would be a very different way of approaching the problem, though there are some insights that might be useful in and of themselves.

Following the publication of Bettencourt et al. (2007) and other, similar work, some useful suggestions were made for studies such as these. Particularly relevant for our application is the requirement to fit models to per capita data (see Figure 19), if that is the quantity of interest to us (Shalizi, 2011). In the same work and in others (Leitão et al., 2016) we can find methods for investigating these types of data – perhaps there exists a statistical model already that would work for us. Also presented are various methods for robust specification and testing of null hypotheses, and discussion around competing models and in which circumstances one might conclude that "scaling is nonlinear" (Leitão et al., 2016).

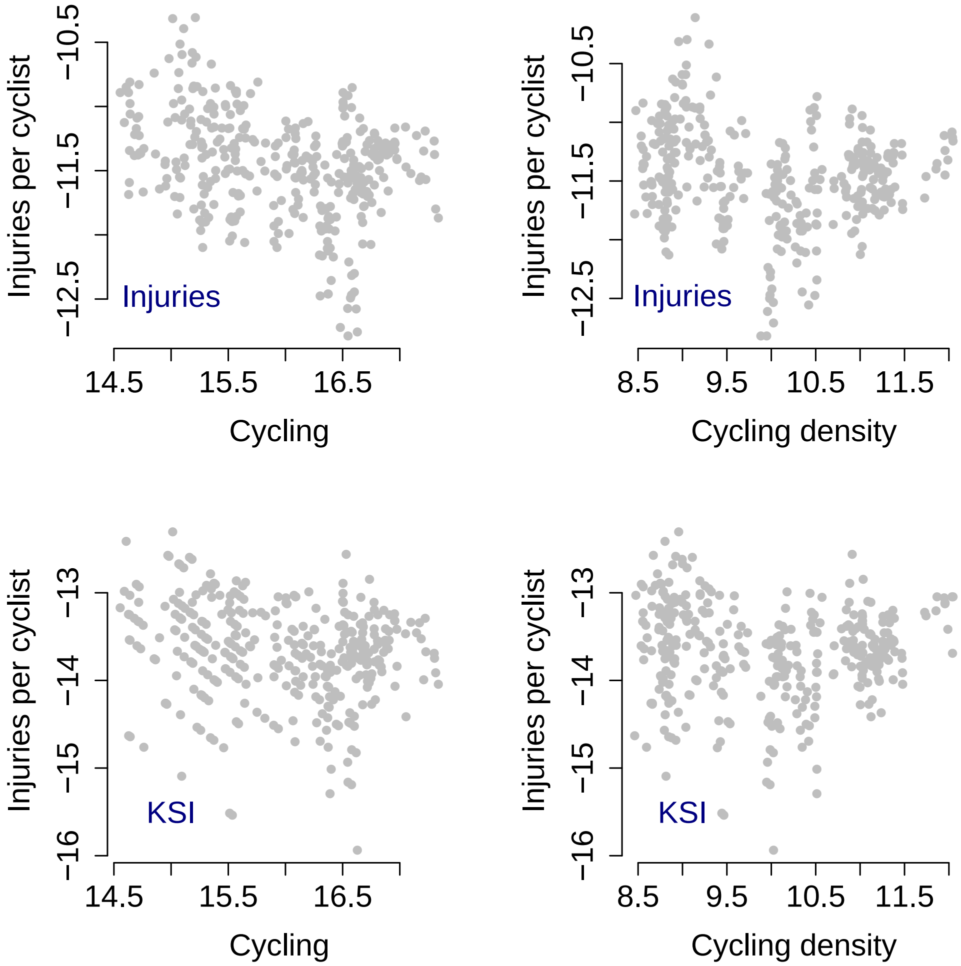

Figure 19: London injuries per cyclist distance, as a function of cyclist distance (left) and cyclist density (right). Cyclist density is calculated as total distance travelled divided by total distance of A, B and minor roads.

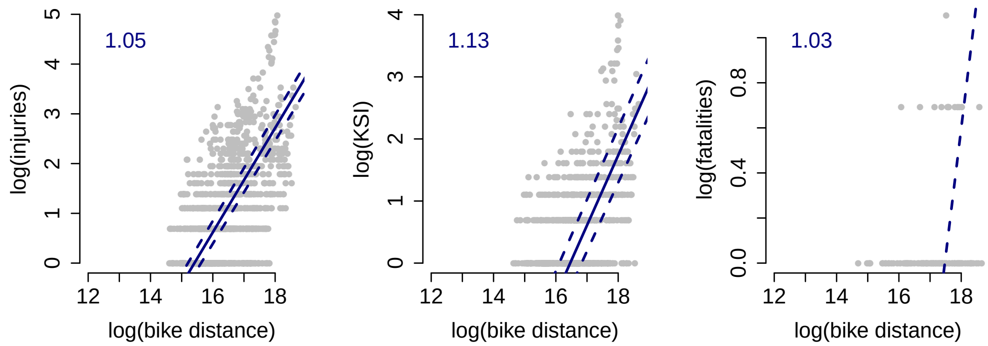

Figure 20: As in Figure 1, counting only injuries that occur in incidents involving no other vehicle (NOV).

Here are some questions, specifically with that field in mind:

(1) Is there a power-law scaling occurring in road injuries? Considering numbers alone (if we want to consider numbers at all), it seems to depend on exactly what we measure and exactly how we model it, e.g. injuries vs. KSI vs. fatalities (see e.g. Table 2). What does that mean for our interpretation? Study of no-other-vehicle casualties might be relevant here (Figure 1 vs. Figure 20).

(2) Within a power-law scaling model, can we elicit the contributions of different modes? Could we apply a latent variable model, and learn the effects of fluctuations of mode types on the outcome?

(3) Can we identify local (protective) effects?

(4) If yes, can we design a mechanism that links local effects to the power-law scaling model, i.e. explains the deviance of $\beta_1 + \beta_2$ from 1 as a function of system size?

References

Aldred, R., Goel, R., Woodcock, J. and Goodman, A. (2017), 'Contextualising Safety in Numbers: a longitudinal investigation into change in cycling safety in Britain, 1991–2001 and 2001–2011', *Injury Prevention* pp. injuryprev–2017–042498.
URL: *http://injuryprevention.bmj.com/lookup/doi/10.1136/injuryprev-2017-042498*

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. and West, G. B. (2007), 'Growth, innovation, scaling, and the pace of life in cities', *PNAS* **104**(17), 7301–7306.

Birrell, P. J., Ketsetzis, G., Gay, N. J., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X.-S., White, P. J., Pebody, R. G. and De Angelis, D. (2011), 'Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London', *Proceedings of the National Academy of Sciences* **108**(45), 18238–18243.
URL: *http://www.pnas.org/cgi/doi/10.1073/pnas.1103002108*

Elvik, R. and Bjørnskau, T. (2017), 'Safety-in-numbers : A systematic review and meta-analysis of evidence', *Safety Science* **92**(0349), 274–282.
URL: *http://dx.doi.org/10.1016/j.ssci.2015.07.017*

Garder, P., Leden, L. and Pulkkinen, U. (1998), 'Measuring the safety effect of raised bicycle crossings using a new research methodology', *Transportation Research Record* **1636**(98), 64–70.
URL: *http://nacto.org/wp-content/uploads/2010/08/Measuring-the-Safety-Effect-of-Raised-Bicycle-Crossings-Using-a-New-Research-Methodology.pdf*

Geyer, J., Raford, N., Pham, T. and Ragland, D. (2006), 'Safety in numbers: data from Oakland, California', *Transportation Research Record: Journal of the Transportation Research Board* **1982**, 150–154.
URL: *http://dx.doi.org/10.3141/1982-20*

Kiers, H. A. and Smilde, A. K. (2007), 'A comparison of various methods for multivariate regression with highly collinear variables', *Statistical Methods and Applications* **16**(2), 193–228.

Leden, L. (2002), 'Pedestrian risk decrease with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario', *Accident Analysis and Prevention* **34**, 457–464.

Leitão, J. C., Miotto, J. M., Gerlach, M. and Altmann, E. G. (2016), 'Is this scaling nonlinear?', *Royal Society open science* **3**, 150649.
URL: *http://dx.doi.org/10.1098/rsos.150649*

Miranda-Moreno, L., Strauss, J. and Morency, P. (2011), 'Disaggregate exposure measures and injury frequency models of cyclist safety at signalized intersections', *Transportation Research Record: Journal of the Transportation Research Board* **2236**, 74–82.
URL: *http://trrjournalonline.trb.org/doi/10.3141/2236-09*

Nordback, K., Marshall, W. E. and Janson, B. N. (2014), 'Bicyclist safety performance functions for a U.S. city', *Accident Analysis and Prevention* **65**, 114–122.
URL: *http://dx.doi.org/10.1016/j.aap.2013.12.016*

Prato, C. G., Kaplan, S., Rasmussen, T. K., Hels, T., Giacomo, C., Kaplan, S. and Rasmussen, T. K. (2016), 'Infrastructure and spatial effects on the frequency of cyclist-motorist collisions in the Copenhagen Region', *Journal of Transportation Safety & Security* **8**(4), 346–360.
URL: *http://dx.doi.org/10.1080/19439962.2015.1055414*

Schepers, J. P. and Heinen, E. (2013), 'How does a modal shift from short car trips to cycling affect road safety?', *Accident Analysis and Prevention* **50**, 1118–1127.
URL: *http://dx.doi.org/10.1016/j.aap.2012.09.004*

Schepers, J. P., Kroeze, P. A., Sweers, W. and Wüst, J. C. (2011), 'Road factors and bicycle – motor vehicle crashes at unsignalized priority intersections', *Accident Analysis and Prevention* **43**(3), 853–861.
URL: *http://dx.doi.org/10.1016/j.aap.2010.11.005*

Shalizi, C. R. (2011), 'Scaling and hierarchy in urban economies', *arXiv* p. 1102.4101v2.

Sim, A., Yaliraki, S. N., Barahona, M. and Stumpf, M. P. H. (2015), 'Great cities look small', *Journal of The Royal Society Interface* **12**(109), 20150315.
URL: *http://rsif.royalsocietypublishing.org/lookup/doi/10.1098/rsif.2015.0315*