Updating and transferring Random Effect models: The case of operating speed percentile estimation

(Article begins on next page)

Transportation Research Part A

Updating and transferring Random Effect models: the case of operating speed percentile estimation

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Paper
Keywords:	Operating speed, speed quantiles (percentile), random effects regression, jack-knife resampling technique, out-of-sample prediction
Corresponding Author:	Cinzia Cirillo University of Maryland College Park, MD United States
First Author:	Jean-Michel Tremblay, PhD
Order of Authors:	Jean-Michel Tremblay, PhD
	Cinzia Cirillo
	Marco Bassani, PhD
Abstract:	Random Effect (RE) models are used for analyzing data that are non-independent or when data are characterized by a hierarchical structure. In traffic and highway engineering, RE models have been successfully employed to estimate free-flow speed distributions from data containing observations that are naturally nested according to different levels (i.e. direction, sections, roads). Empirical studies conducted on both urban arterials and rural two-lane highways have shown that RE models, by properly accounting for the survey design, are superior to traditional Fixed Effect (FE) models. In this paper, the transferability of RE models to road sections that were not in the original sample used for model estimation was studied, under the assumption that for these additional sections very few observations are available or can be collected. This problem poses two challenges. First, random effects for the new road sections should be estimated in order to make out-of-sample predictions. Second, the original model formulation makes use of speed quantiles as predictors of the linear model which are not readily available for the new sections. The method proposed estimates an auxiliary model, in which the RE of the original model are correlated to the RE to be defined for the new section, with the former being used to predict the latter. The RE pairs are modeled jointly, taking advantage of their potential mutual correlation. The model coefficients obtained are also validated using a jackknife technique. Results show that the method converges quite fast and that a handful of observations for the new road section are sufficient for good RE estimates.
Suggested Reviewers:	
Opposed Reviewers:	

Updating and transferring Random Effect models: the case of operating speed percentile estimation

Jean-Michel Tremblay[1], Cinzia Cirillo[2], and Marco Bassani[3]

[1]*Edgybees Ltd., jeanmi.tremblay@gmail.com*
[2]*University of Maryland, Department of Civil and Environmental Engineering, ccirillo@umd.edu*
[3]*Politecnico di Torino, Department of Environment, Land and Infrastructure Engineering, marco.bassani@polito.it*

April 20, 2020

Abstract

Random Effect (RE) models are used for analyzing data that are non-independent or when data are characterized by a hierarchical structure. In traffic and highway engineering, RE models have been successfully employed to estimate free-flow speed distributions from data containing observations that are naturally nested according to different levels (i.e. direction, sections, roads). Empirical studies conducted on both urban arterials and rural two-lane highways have shown that RE models, by properly accounting for the survey design, are superior to traditional Fixed Effect (FE) models.

In this paper, the transferability of RE models to road sections that were not in the original sample used for model estimation was studied, under the assumption that for these additional sections very few observations are available or can be collected. This problem poses two challenges. First, random effects for the new road sections should be estimated in order to make out-of-sample predictions. Second, the original model formulation makes use of speed quantiles as predictors of the linear model which are not readily available for the new sections.

The method proposed estimates an auxiliary model, in which the RE of the original model are correlated to the RE to be defined for the new section, with the former being used to predict the latter. The RE pairs are modeled jointly, taking advantage of their potential mutual correlation. The model coefficients obtained are also validated using a jackknife technique. Results show that the method converges quite fast and that a handful of observations for the new road section are sufficient for good RE estimates.

Keywords: Operating speed, speed quantiles (percentile), random effects regression, jackknife resampling technique, out-of-sample prediction. .

1 Introduction

Among the parameters characterizing vehicular flows, speed is used in multiple applications including traffic analysis, speed management, road design, and road safety. Speeds are collected through spot speed observations or by recurring to permanent acquisition units in the field (Garber and Hoel 2020; Catani et al. 2017). This data may be used to calibrate models to explain how speed is related to some significant variables depicting the road scenario. As a result, analysts and road designers can select the most appropriate road features to modulate risk perception and compel drivers to adopt consistent speed decisions and behaviors. Starting from the ‘80s, a conspicuous quantity of papers proposed models to predict the 85th speed percentile (i.e., V85) of operating speeds (OS), and OS differential (e.g., ΔV85, 85MSR) between road elements (Dimaiuta et al. 2011).

V85 is conventionally considered representative of OS distribution since it separates the speed of prudent drivers from that of more aggressive ones. V85 models may include geometric characteristics of roads (e.g., lane width, radius or curvature), environmental conditions (e.g., lighting, weather, land use), and driving regulations affecting driver behavior (Himes et al. 2013). The variation in operating speed between road elements has repeatedly been used to support design decisions (Lamm et al. 1988). Park and Saccomanno (2006) evidenced that speed variations must be evaluated for individual drivers (i.e., disaggregated data) rather than from aggregated data for the observed group, in order to prevent the so called “ecological fallacy” problem. However, this approach is challenging due to the need to monitor individual vehicles along entire road segments (McFadden and Elefteriadou 2000).

However, when attention is focused on a section, the use of V85 becomes controversial since different OS distributions may exhibit the same 85th percentile. To address this issue, Shankar and Mannering (1998) proposed the use of simultaneous equations to model the average and standard deviation of speed in each lane of multilane highways. Later on, Figueroa-Medina and Tarko 2005 introduced a model to predict any speed percentile combining the mean and the standard deviation in a linear regression equation.

Most of the available literature has proposed models of the Fixed Effect (FE) type, in which each speed observation along a road section is assumed to be dependent on the predictors included in that model only (Dimaiuta et al. 2011). This is only acceptable when speed clusters used to calibrate the model are not distinguished per direction, are from segments that do not belong to the same road, and are sufficiently distant from each other. However, when speed observations are clustered and spatially close to each other, each of them may share unobserved effects. Thus, it is not possible to assume independence of errors for individual observations without considering Random Effects (RE) for these groups.

Tarris et al. (1996) carried out a panel analysis of free-flow speed data collected from individual drivers with speed values recorded at sensor locations. In the proposed model, RE were associated with groups and speed location. Islam and El-Basyouny (2015) used RE to account for differences in hourly free-flow speed data related to site and community in a pilot study aimed at reducing OS. More recently, Cheng et al. (2018) evidenced that RE are fundamental in predicting the speed and speed deviation along lanes in multilane highways. They used RE to account for the variation between adjoining lanes, between adjacent segments, and among segments.

RE models reach a higher coefficient of determination than those obtained assuming FE coefficients for groups and sensor locations (Bassani, Dalmazzo, et al. 2014). RE models for OS was proposed by the authors (Bassani, Dalmazzo, et al. 2014; Bassani, Cirillo, et al. 2016; Bassani, Catani, et al. 2016) in multiple observations for the same direction (d) of a section (s), on several sections of a road (r), and on several roads of the network. In the model:

$$V_{rsd,i} = \beta_0 + \beta_k X_{rsd,k} + \beta_j Z_p X_{rsd,j} + \alpha_r + \alpha_{s|r} + \alpha_{d|rs} + \epsilon_{rsd,i}, \tag{1}$$

where β_0 is the general model intercept, β_k and β_j are calibration parameters for the k and $j$variables affecting the estimated mean $X_{rsd,k}$, and the estimated standard deviation $X_{rsd,j}$ respectively, and Z_p is the standardized normal variable. In eq. 1, α_r, $\alpha_{s|r}$ and $\alpha_{d|rs}$ are the three nested RE accounting for the variability introduced by the random selection of roads in the network, the section within a road, and the directions in the section.

The objective of this study is to transfer RE models calibrated on a given sample of lanes, sections and roads to road sections that were not included in the estimation sample and for which few speed observations were collected or available to the analyst. The problem has relevant practical implications, as the method proposed will facilitate the use of existing models on different sections without the need to collect a significant number of new observations, which is usually a lengthy and costly process. In order for this transferability to be effective and to have realistic out-of-sample predictions, RE need to be predicted for the new road section. Also, the model specification is based on speed quantiles, which are not readily available for the new road sections.

REs in most situations are assumed to have zero mean and therefore the best a priori predictor for REs in a new road section is zero as well. However, it is likely that better predictors can be produced by considering a simpler, auxiliary RE model whose purpose is to overcome the unavailability of quantiles in the validation sample. Specifically, following McCulloch et al. (2008), it is possible to build an auxiliary model which, although inferior to the actual model, takes advantage of the potential correlation existing across RE pairs in order to provide good RE predictors. The method is based on the assumption that REs of the original model are correlated with the REs to be defined for the new section and that the first one can be used to predict the latter. The methodology is first developed in the case of one RE, and then extended to the case of a model including two REs. The best predictor is derived and the convergence is tested on empirical data. Finally, validation is performed on the quantiles of all the road sections considered using a jackknife technique (Efron and Tibshirani 1993).

The remaining of this paper is articulated as follows. Model formulation for one RE and two REs is presented in Section 2. Numerical results derived from real data collected in the North-West of Italy are reported in Section 3. The case where predictors are multiplied by the normal quantiles is solved in Section 4. In Section 5, a Jackknife re-sampling technique is used to analyze the causes of poor estimation results. Conclusions and suggestions for future research are given in Section 6.

2 Model formulation

In this Section, we derive the statistical method to transfer an estimated random effect model to a new road section for which very few observations are available to the analyst. The problem is that for this new road section, only the model predictors are available, while the random effect(s) are unknown. Under the hypothesis that the random effect of the new section are correlated to those of the section for which the model has been estimated, we derive the conditional mean of the *Best Linear Unbiased Predictor* (BLUP) of the error in the new section. The method is developed first for a one random effect model (Section 2.1) and then generalized to a two random effect model (Section 2.2).

2.1 One random effect model

Following the formulation in eq.1, we first develop the proposed methodology for a simple case that contains one random effect. Let's assume that speed data is available for several sections s, and that the response variable $V_{s,i}$ is affected by a set of predictors $X_{s,k}$. The model contains one random effect α_s and an error term $\epsilon_{s,i}$:

$$V_{s,i} = \beta_0 + \beta_k X_{s,k} + \alpha_s + \epsilon_{s,i}, \tag{2}$$

where α_s and $\epsilon_{s,i}$ are independent and follow a normal distribution:

$$\begin{aligned} \alpha_s &\sim N(0, \sigma_s^2), \\ \epsilon_{s,i} &\sim N(0, \sigma^2). \end{aligned}$$

while β is defined as the the vector of fixed coefficients to be estimated and that includes both β_0 and β_k.

Once the model is calibrated, estimates for the model parameters (denoted as $\hat{\beta}$, $\hat{\sigma}_s^2$, and $\hat{\sigma}^2$) are available to the analyst, and predictions for the random variables α_s and $V_{s,i}$ can be obtained (McCulloch et al. 2008).

Assuming the analyst is interested in predicting random effects for a new section s' with unknown random effect $\alpha_{s'}$, the best *a priori* estimate of $\alpha_{s'}$ is zero since $\alpha_{s'} \sim N(0, \sigma_s^2)$. In some contexts this may be satisfying, but other methods can be explored under the assumption that it is possible to sample a few observations for the new section s' in order to improve the knowledge about $\alpha_{s'}$ and build a better *a posteriori* prediction. This is precisely the ultimate objective of our method.

Suppose that n observations $V_{s',1}, V_{s',2}, \ldots, V_{s',n}$ are collected in the new section. $\alpha_{s'}$ cannot be observed directly, because we always observe the *sum* of the errors. Hence according to eq. 2 the sum of the residuals is:

$$r_{s',i} = \alpha_{s'} + \epsilon_{s',i} = V_{s',i} - \beta X_{s',k}.$$

The difference $V_{s',i} - \beta X_{s',k}$ cannot be measured because the value of the parameters in β is unknown, so the estimated value $\hat{\beta}$ from model calibration is used to approximate the total residuals r'_s, i.

Let's define $u = (\alpha_{s'})$ and let's assume that it follows a normal distribution with mean zero and variance $D = (\sigma_s^2)$. In this case, u refers to a single random effect, but as we will see in the next Section, the methodology applies to any number of effects to be predicted. The residuals $r = (r_{s',1}, r_{s',2}, ..., r_{s',n})$ have zero mean and variance W:

$$W = \begin{bmatrix} \sigma_s^2 + \sigma^2 & \sigma_s^2 & ... & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma^2 & ... & \sigma_s^2 \\ ... & ... & ... & ... \\ \sigma_s^2 & \sigma_s^2 & ... & \sigma_s^2 + \sigma^2 \end{bmatrix}.$$

W can be written as follows:

$$W = \sigma^2 I + \sigma_s^2$$

where I is the identity matrix.

The covariance between the random effect $\alpha_{s'}$ and one given total residual is:

$$Cov(\alpha_{s'}, r_{s',i}) = \sigma_s^2.$$

and more generally, the covariance C between u and r is:

$$C = \begin{bmatrix} \sigma_s^2 & \sigma_s^2 & ... & \sigma_s^2 \end{bmatrix} = \sigma_s^2 1_{1\times n}$$

and therefore:

$$\begin{bmatrix} u \\ r \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D & C \\ C^T & W \end{bmatrix} \right).$$

At this stage, it is not necessary to formulate hypotheses on the joint distribution of (u, r); it is sufficient to know the first and second moments of (u, r) to derive the *best linear unbiased predictor* (BLUP) of u (McCulloch et al. 2008). If the joint distribution were normal, the BLUP would be the overall best predictor of u. This not the case for our real application.

Once r is observed, the BLUP of u is the conditional mean. The expectation of $u|r$ is given by:

$$\hat{\alpha}_{s'} = E(u|r) = CW^{-1}r.$$

The estimated values of the coefficients are used to produce numerical values for the BLUP. The predicted value for an observation i in section s' will be:

$$\hat{V}_{s',i} = \hat{\beta}X_{s',k} + \hat{\alpha}_{s'}.$$

$\hat{\beta}$ can be estimated in a relatively easy way, so the main challenge for this problem is to predict $\hat{\alpha}_{s'}$. The objective here is to investigate what is the smallest sample in section s' that we can use to predict $\alpha_{s'}$ satisfactorily. In general, it can be observed that the prediction converges faster when σ^2 is low relative to σ_s^2; this is because one single observation of r is expected to be less noisy and more correlated with the unknown realized value $\alpha_{s'}$.

2.2 Two random effects models

The case with two nested random effects is formulated in eq. 3:

$$V_{sd,i} = \beta_0 + \beta_k X_{sd,k} + \alpha_s + \alpha_{d|s} + \epsilon_{sd,i}, \tag{3}$$

where the subscript d stands for direction, so speed data can be distinguished into two different directions for the same section. The random effect $\alpha_{d|s}$ is nested within the levels of α_s.

We make the assumption that this new random effect is normally distributed:

$$\alpha_{d|s} \sim N(0, \sigma_d^2)$$

As a result, in this model a sample from the new section s' also includes some levels of the direction effect. For illustration purposes, one section, two directions and n observations per direction are assumed. The RE to be predicted is:

$$u = (\alpha_{s'}, \alpha_{d1}, \alpha_{d2}),$$

and the total residuals are given by:

$$r_{s'd,i} = V_{s'd,i} - \beta X_{s'd,k}.$$

where β is the vector of coefficient to be estimated.

Similar to the one RE model, the RE cannot be directly observed. In this case, the covariance structure for the total residuals is:

$$\begin{aligned} var(r_{s'd,i}) &= \sigma_s^2 + \sigma_d^2 + \sigma^2, \\ cov(r_{s'd,i}, r_{s'd,j}) &= \sigma_s^2 + \sigma_d^2, \\ cov(r_{s'd,i}, r_{s'd',i}) &= \sigma_s^2 \end{aligned}$$

Two residuals in the same section but for different directions only share the α_s term so their covariance is σ_s^2. Residuals in the same direction also share $\alpha_{d|s}$ so their covariance is the sum $\sigma_s^2 + \sigma_d^2$. Finally, the total variance of r is the sum of its three components $\sigma_s^2 + \sigma_d^2 + \sigma^2$. These results are a consequence of the independence assumption made on the three REs.

As in the previous section, we want to derive the joint moments of (u, r). The covariance matrices D, C and W of (u, r) are respectively:

$$D = \begin{bmatrix} \sigma_s^2 & 0 & 0 \\ 0 & \sigma_d^2 & 0 \\ 0 & 0 & \sigma_d^2 \end{bmatrix}, W = \sigma_s^2 + \begin{bmatrix} \sigma_d^2 + \sigma^2 & \sigma_d^2 & \sigma_d^2 & 0 & 0 & 0 \\ \sigma_d^2 & \sigma_d^2 + \sigma^2 & \sigma_d^2 & 0 & 0 & 0 \\ \sigma_d^2 & \sigma_d^2 & \sigma_d^2 + \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_d^2 + \sigma^2 & \sigma_d^2 & \sigma_d^2 \\ 0 & 0 & 0 & \sigma_d^2 & \sigma_d^2 + \sigma^2 & \sigma_d^2 \\ 0 & 0 & 0 & \sigma_d^2 & \sigma_d^2 & \sigma_d^2 + \sigma^2 \end{bmatrix}$$

with $\mathbf{0}$ being a matrix of zeros and I the identity matrix:

$$W = \sigma_s^2 + \begin{bmatrix} \sigma_d^2 + \sigma^2 I & 0_{n\times n} \\ 0_{n\times n} & \sigma_d^2 + \sigma^2 I \end{bmatrix}$$

C is finally given by:

$$C = \begin{bmatrix} \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_d^2 & \sigma_d^2 & \sigma_d^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_d^2 & \sigma_d^2 & \sigma_d^2 \end{bmatrix} = \begin{bmatrix} \sigma_s^2 1_{1\times n} & \sigma_s^2 1_{1\times n} \\ \sigma_d^2 1_{1\times n} & 0_{1\times n} \\ 0_{1\times n} & \sigma_d^2 1_{1\times n} \end{bmatrix}$$

The predicted random effects are given by the expected mean:

$$E(u|r) = CW^{-1}r.$$

The predicted value for an observation i in section s' and direction d will be:

$$\hat{V}_{s'd,i} = \hat{\beta}_{s'di} X_{s'd,k} + \hat{\alpha}_{s'} + \hat{\alpha}_{d|s}.$$

3 Numerical examples on OS data

Speed data used in this study were collected in road sections of two-lane rural highways in the North-West part of Italy. Individual speeds of isolated vehicles were collected under free-flow conditions in sections where vehicles travel at constant speed (i.e. in the center of tangents and curves). Speeds were included in the database only when a minimum headway of six seconds was observed. The data for model estimation were extracted from a larger database already used by the authors in (Bassani, Cirillo, et al. 2016). The density and the presence of elements along the road section was evaluated along one km across the sample sections. Table 1 lists the values assumed by the variables that were found to be significant in the calibration of the model reported in eq. 1. In the table, the variables are divided into those affecting the average (X_k) and the dispersion (X_j) of predictors. The latter are also divided into numerical and Boolean variables. Finally, the last five columns summarize the minimum (V_{min}), the maximum (V_{max}), the 50$^{\text{th}}$ ($V50$) and the 85$^{\text{th}}$ ($V85$) percentile of speeds included in the database, while n_{obs} indicates the number of data available for each section. The notes at the end of table 1 describe the acronyms used to identify the variables.

Table 1: Summary of characteristics of the selected road sections

Sections	Significant variables in model 1																			Speed data				
	X_{ki}		X_{ji} (numerical)											X_{ji} (boolean)										
	1/R	PedD	LW	SLW	SRW	LG	TRDLS	TRDRS	DDRS	IDLS	IDRS	PSL	ΔPSL	LBLS	LRRS	SLS	SRS	SBLS	SBRS	V_{min}	V_{50}	V_{85}	V_{max}	N
	m^{-1}	n/km	m	m	m	%	n/km	n/km	n/km	n/km	n/km	km/h	km/h	-	-	-	-	-	-	km/h	km/h	km/h	km/h	n
1	0.00	0	3.60	0.40	0.40	1.50	0	0	2	1	0	70	0	1	1	1	0	0	0	39.0	66.0	76.0	97.0	429
3	0.00	0	3.75	0.40	0.40	0.00	0	1	2	0	0	50	0	1	1	0	0	1	1	32.0	74.0	85.0	157.0	972
5	0.00	0	3.70	0.00	0.00	1.00	0	0	1	0	0	90	0	0	1	0	0	1	0	38.0	74.0	89.0	130.0	669
7	3.29e−3	0	3.75	1.50	1.50	5.14	0	0	0	0	0	70	0	1	1	0	0	1	1	24.0	82.0	94.0	114.0	312
8	0.00	0	3.75	1.50	1.50	2.09	0	0	0	0	0	70	0	1	1	0	0	1	1	57.0	86.0	98.0	124.0	101
9	3.10e−4	0	3.75	1.50	1.50	4.69	0	0	0	1	0	70	0	0	1	0	0	1	0	52.0	80.0	91.0	114.0	101
10	6.67e−3	0	3.25	0.00	0.00	8.50	0	1	3	0	2	50	0	0	0	0	0	1	1	46.0	62.0	70.1	77.0	87
11	0.00	0	3.00	0.50	0.50	1.50	0	0	4	0	3	70	0	0	0	0	0	0	0	42.6	71.0	90.3	129.4	107
12	0.00	0	3.00	0.50	0.50	0.00	0	0	1	0	1	70	0	0	0	0	0	0	0	46.7	81.5	94.3	128.6	120
13	0.00	0	3.00	0.50	0.50	0.00	0	0	1	0	0	70	0	0	0	0	0	0	0	50.0	80.2	90.4	114.0	108
14	0.00	0	3.00	0.50	0.50	2.00	0	0	2	0	0	70	0	0	0	0	0	0	0	49.6	83.0	101.7	129.4	125
15	0.00	0	3.00	0.50	0.50	0.00	0	0	0	0	1	70	0	0	0	0	0	0	0	49.0	82.7	95.9	127.8	127
16	0.00	0	3.00	0.50	0.50	0.00	0	0	1	0	0	70	0	0	0	0	0	0	0	46.3	83.2	98.9	128.3	138
17	0.00	0	3.00	0.50	0.50	0.00	0	0	6	2	1	70	0	0	0	0	0	0	0	49.4	78.5	95.3	149.6	128
18	0.00	0	3.50	1.20	1.20	0.50	0	0	1	0	0	90	0	0	0	0	0	0	0	42.0	67.2	88.5	115.2	41
20	4.44e−4	0	3.50	1.20	1.20	0.50	0	0	0	1	1	90	0	0	0	0	0	0	0	70.0	97.0	107.0	118.0	26
21	0.00	0	3.50	1.20	1.20	0.50	1	0	0	4	4	90	0	0	0	0	0	0	0	58.0	99.5	112.9	127.0	28
22	0.00	0	3.50	1.20	1.20	0.00	0	0	5	1	1	90	0	0	1	0	0	0	0	55.0	77.0	100.3	130.0	30
23	6.66e−4	0	3.50	1.20	1.20	1.00	0	0	1	1	4	50	0	0	0	0	0	1	1	32.0	50.0	57.4	72.0	32
24	0.00	0	3.50	1.20	1.20	0.00	0	0	8	0	3	70	0	1	0	0	1	0	0	59.0	77.0	92.6	110.0	27
25	1.10e−3	0	3.50	1.20	1.20	1.00	0	0	1	0	0	70	0	0	0	0	0	1	1	44.0	69.0	78.0	96.0	37
26	1.20e−4	0	3.50	1.20	1.20	0.00	0	0	6	0	2	70	0	1	0	0	1	0	0	47.0	75.0	85.5	98.0	38
27	2.21e−3	4	3.50	1.20	1.20	0.00	0	0	4	3	3	50	0	0	1	1	1	0	0	44.0	51.0	57.5	70.0	36
28	6.65e−4	0	3.50	1.20	1.20	1.00	0	0	2	0	0	90	0	0	0	0	0	0	0	56.0	87.0	97.0	130.0	43
29	0.00	0	3.50	1.20	1.20	1.00	1	0	1	0	0	90	0	0	1	0	0	0	0	65.0	91.0	107.9	119.5	38
30	0.00	1	3.50	1.20	1.20	0.00	0	0	7	3	5	70	0	0	0	1	1	0	0	54.7	69.3	85.1	108.1	42
31	0.00	0	3.50	1.20	1.20	0.00	0	0	6	0	1	70	0	1	0	0	1	0	0	44.3	73.1	83.6	120.2	140
32	3.33e−3	2	3.20	0.80	0.80	0.00	0	0	4	4	2	70	0	0	1	0	0	0	0	27.6	66.7	75.8	93.3	116
33	0.00	0	3.80	1.26	1.26	0.00	0	0	3	1	0	90	0	0	0	0	1	0	0	42.6	67.4	75.3	106.6	67
34	0.00	3	3.50	0.40	0.40	0.00	0	0	5	1	4	50	0	0	1	1	1	0	0	34.7	55.0	68.5	105.4	154
36	0.00	1	3.50	0.70	0.70	0.00	0	0	8	1	1	50	0	1	1	0	0	0	0	36.9	69.1	82.5	134.8	161

Notes: 1/R = curvature, PedD = pedestrian density, LW = lane width, SLW = shoulder left width, SRW = shoulder right width, LG = longitudinal grade, TRDLS = total ramp density left side, TRDRS = total ramp density right side, DDRS = driveway density right side, IDLS = intersection density left side, IDRS = intersection density right side, LBLS = lay-by left side, LBRS = lay-by right side, SLS = sidewalks left side, SLS = sidewalks right side, SBLS = safety barriers left side, SBRS = safety barriers right side, PSL = posted speed limit, Vmin = minimum speed value, V50 = 50[th] percentile of speed, V85 = 85[th] percentile of speed, Vmax = maximum speed value, N = number of speed data. Binomial values = 1 when present, = 0 otherwise. LG is in absolute value and is specified per direction

3.1 Results: one RE model

The first case study models speed data with only one RE, as formulated in 2.1; specifically, it accounts for the RE related to sections but ignores the RE related to directions. A model is calibrated using thirty sections out of the thirty one available in the database and listed in Table 1. Once the model is fitted on the thirty sections, the RE effect is predicted for the section that was left out from model estimation. This is repeated for all the thirty one sections in the dataset. This procedure provides a series of predicted REs; these results are used to assess the convergence of the method proposed.

Figures 1 and 2 illustrate the results obtained for each section. Each subplot represents the predicted RE in the section used for validation. The x-axis corresponds to the number of observations that were used from the validation section in order to predict the REs. For example, an x-value of 10 for section 3 means that a model with all sections but the third one was estimated, and that ten observations in the third section were used in order to predict the (realized) effect of section 3.

Each subplot contains a solid line, a dashed line, and a dotted line. The solid line shows the predicted effects obtained with the **full** model, for which the quantile information in the validation sample is assumed to be known. This is ultimately the effect that we aim to predict. The dotted line shows the predicted effects obtained with a **simple** auxiliary model that only contains the predictors $X_{s,k}$ that affect the mean speed. The dashed line shows the predicted effects for the **conditional** model obtained using information from the auxiliary model. This is the prediction that ultimately is going to be used.

From figures 1 and 2, the following three remarks can be made: (i) a relative convergence in the predicted effects as the number of observations grows is observed; (ii) there is a substantial difference between the dotted line and the black solid line; meaning that the closer the two lines are, the more likely that the one can be predicted from the other; (iii) the dashed line does not approximate the solid line and it is mostly super-posed to the dotted line. Therefore, it can be concluded that the predicted effect of the full model using the auxiliary model are no better approximation than just the predicted effects of the auxiliary model. This results might be seen as disappointing, but later it will be proved that correctly accounting for the sample design solves the problem observed.

Figure 1: Sections 1-22

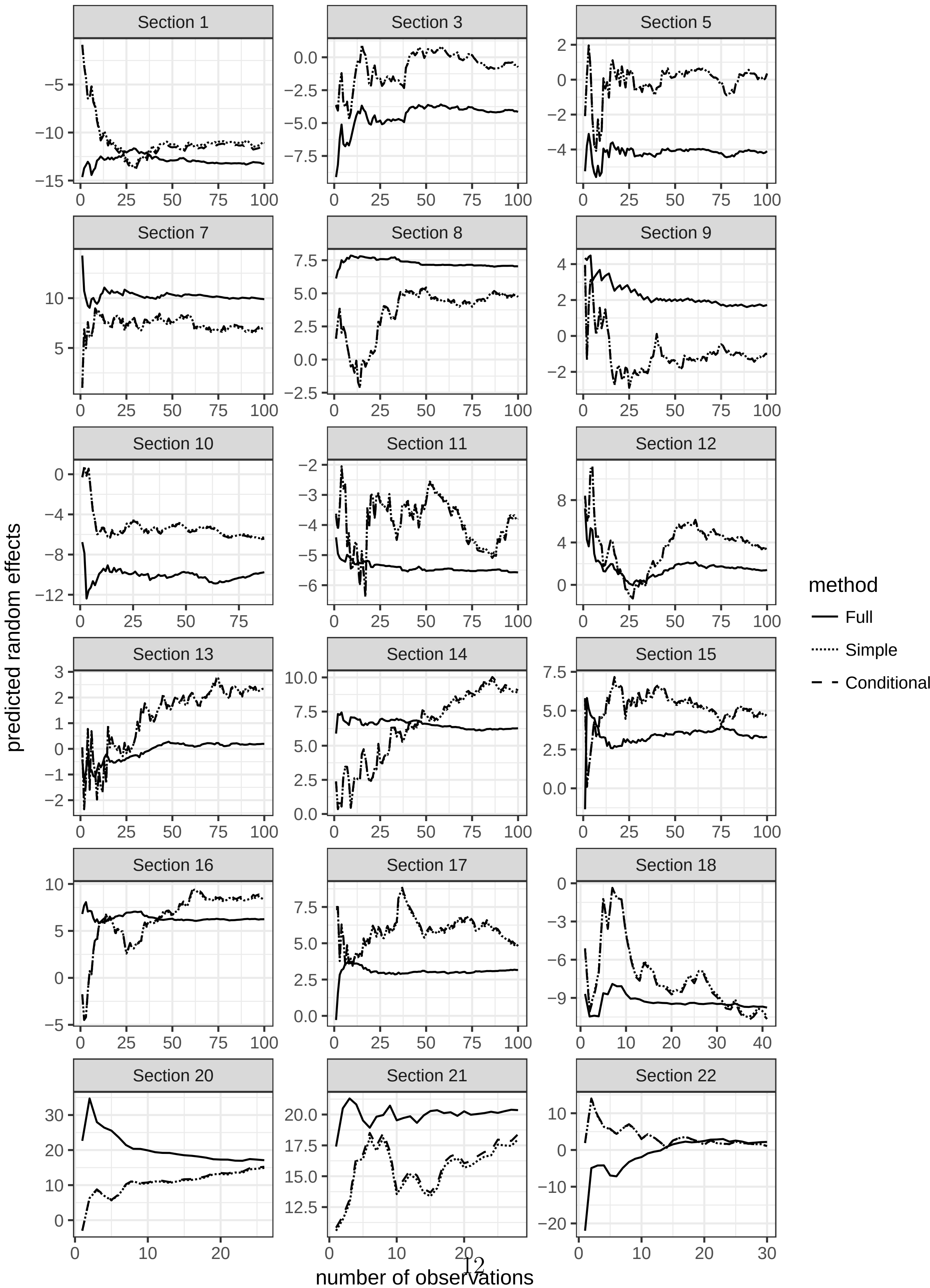
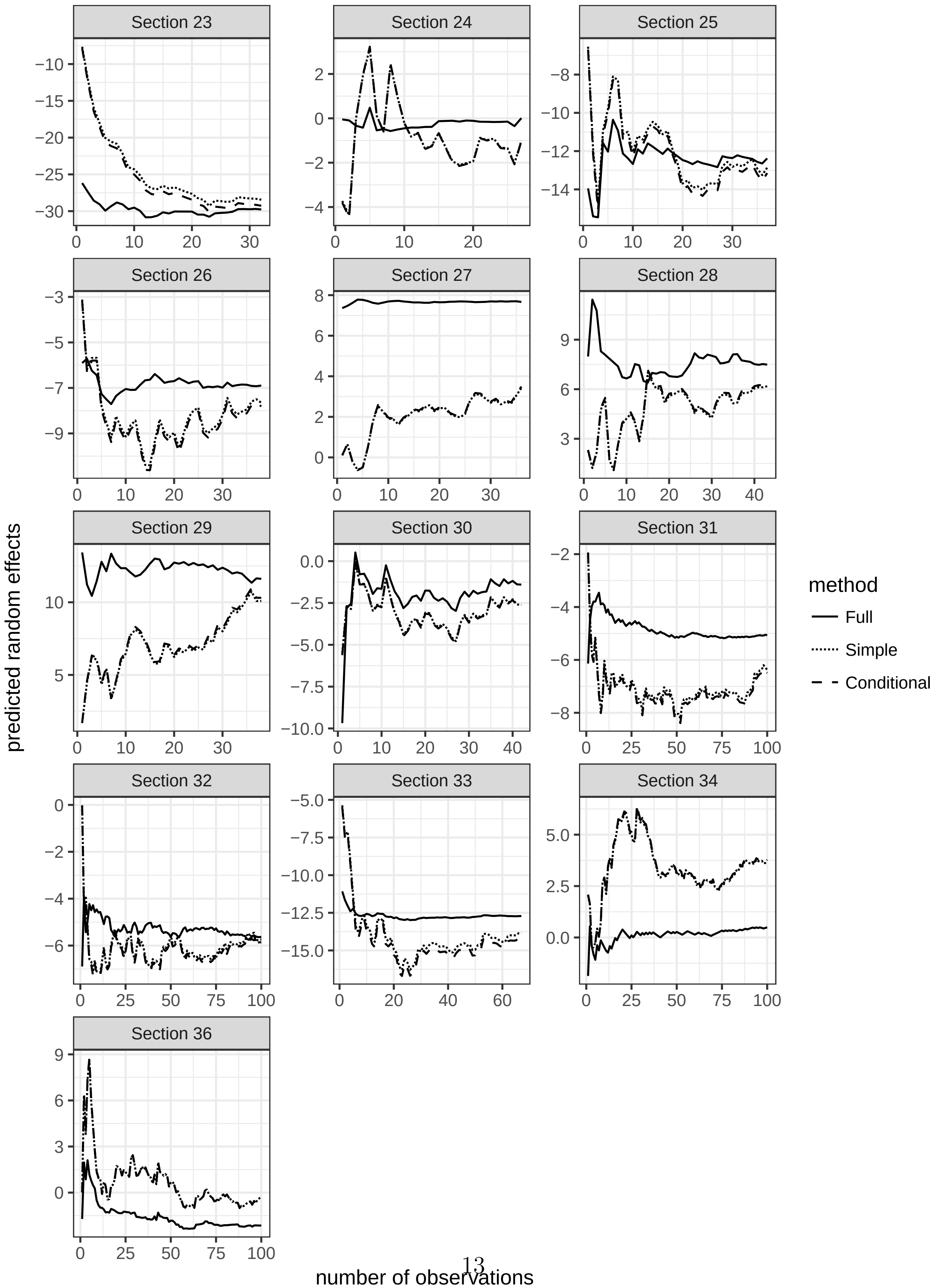


Figure 2: Sections 23-36



3.2 Results: two RE model

The second case study models two REs as formulated in 2.2; this time both section and direction effects are taken into account. The scope here is to assess the convergence of $\hat{\alpha}_{s'}$, $\hat{\alpha}_{d1}$ and $\hat{\alpha}_{d2}$ in the presence of two REs; this model formulation is fully consistent with the survey design.

Figures 3 and 4 should be read in the same way as figures 1 and 2. The thick solid line corresponds to the predicted section effect, and the two fine solid lines correspond to the predicted direction effects. The dashed lines correspond to the predicted effects using the auxiliary model and those are used to compute the REs in the validation section.

For both sections and directions, the predicted effects using the auxiliary model are very close to the ones using the full model. However, there are still noticeable differences when very few observations are used for the prediction. For example, predictions in sections 8 and 27 are not very precise for only two or three observations.

It is worth noting that striking differences exist between predictions with one and two REs. Effects with only the section component are not close to the *true* effects of the full model; incorporating the direction effects, thus accounting for the design of the sample, drastically improves the predictions.

4 Computation of the residuals

We now turn our attention to the case where the model includes variables $X_{sd,j}$ multiplied by the normal quantile Z_p in addition to the regular predictors $X_{sd,k}$, (see Bassani, Dalmazzo, et al. 2014; Bassani, Cirillo, et al. 2016; Bassani, Catani, et al. 2016). This is done to calculate any percentile speed as a linear combination of variables affecting both the central tendency and the dispersion of the collected speed data. Those quantiles are not available for the validation sample, which is too small to allow for the calculation of variance indicators. Note that $Z_{sd,j}$ is a scalar so the term $Z_p X_{sd,j}$ is simply a scalar multiplied by a vector.

$$V_{sd,i} = \beta_0 + \beta_k X_{sd,k} + \beta_j Z_p X_{sd,j} + \alpha_s + \alpha_{d|s} + \epsilon_{sd,i}$$

In this case, it is not possible to isolate the sum of REs and we must rely on alternative methods. The strategy that is proposed here makes use of the estimated total residuals from a restricted model that only contains observable variables ($X_{sd,k}$) to predict the REs of an unrestricted model (with both $X_{sd,k}$ and $X_{sd,j}$).

Figure 3: Sections 1-22

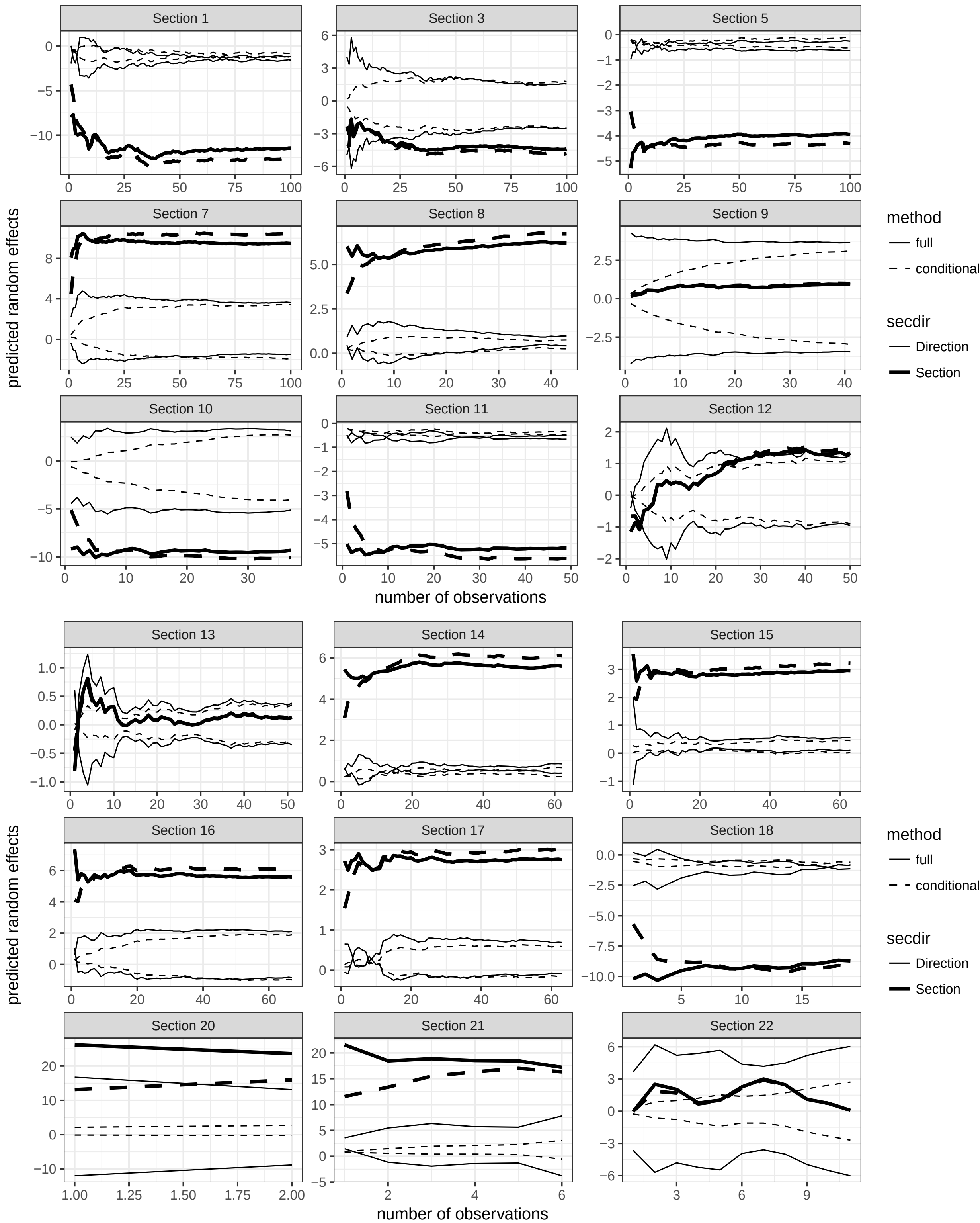
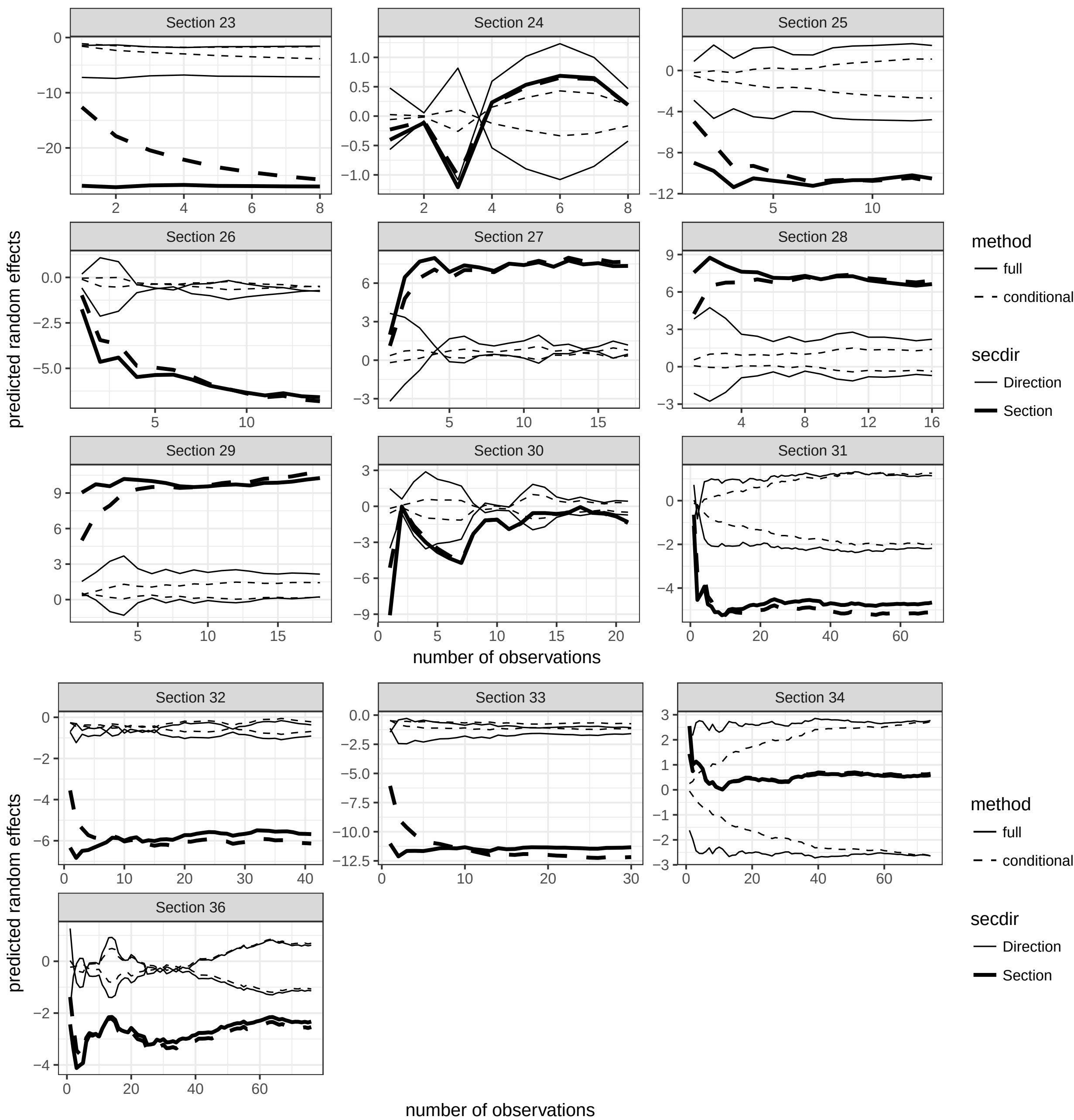


Figure 4: Sections 23-36



$$
\begin{aligned}
V_{sd,i} &= \beta_k X_{sd,k} + \alpha^*_s + \alpha^*_{d|s} + \epsilon^*_{sdi}, \\
\alpha^*_s &\sim N(0, \sigma^{*2}_s), \\
\alpha^*_{d|s} &\sim N(0, \sigma^{*2}_d), \\
\epsilon^*_{sd,i} &\sim N(0, \sigma^{*2}). \\
r^*_{sd,i} &= V_{sd,i} - \beta_k X_{sd,k}
\end{aligned}
\tag{4}
$$

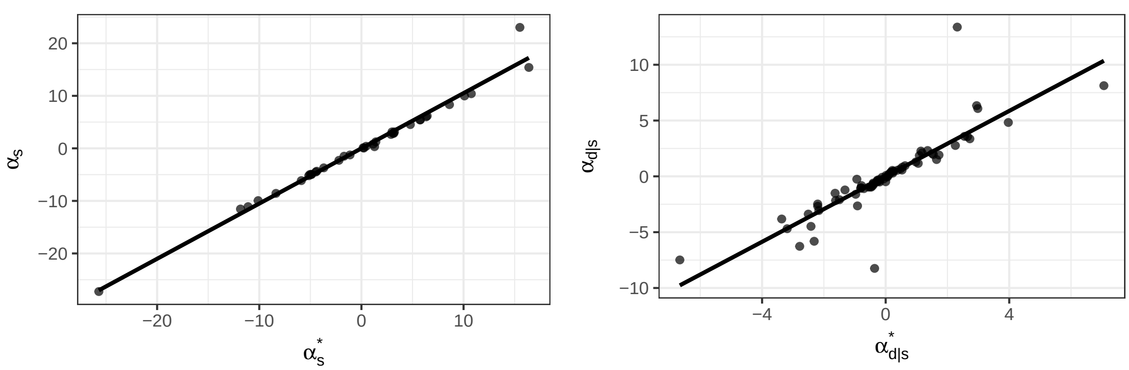
$$
\begin{aligned}
V_{sd,i} &= \beta_k X_{sd,k} + \beta_j X_{sd,j} + \alpha_s + \alpha_{d|s} + \epsilon_{sdi}, \\
\alpha_s &\sim N(0, \sigma^2_s), \\
\alpha_{d|s} &\sim N(0, \sigma^2_d), \\
\epsilon_{sd,i} &\sim N(0, \sigma^2). \\
r_{sd,i} &= V_{sd,i} - \beta_k X_{sd,k} - \beta_j X_{sd,j}
\end{aligned}
\tag{5}
$$

The restricted (eq. 4) and unrestricted models (eq. 5) are calibrated and parameters are estimated, together with the empirical correlation (ρ_{α_s} and $\rho_{\alpha_{d|s}}$,) across the effects of both models. Figures 5a and 5b plot the predicted section and direction effects of the restricted model and the unrestricted model respectively. As anticipated, the more correlated these effects are, the easier will be to predict one using the other.

Figure 5: Predicted effects

(a) Comparison of section effects

(b) Comparison of direction effects



The term r^*_{sdi} can be calculated in the validation sample, while the variance of one of those residuals is given by:

$$Var(r^*_{sd,i}) = \sigma^{*2}_s + \sigma^{*2}_d + \sigma^{*2}.$$

The covariance of two residuals is given by:

$$
\begin{aligned}
Cov(r_{sdi*}, r_{sdj*}) &= \sigma^2_{s*} + \sigma^2_{d*} && \text{in the same direction,, and} \\
Cov(r_{sd,i*}, r_{sd'i*}) &= \sigma^2_{s*} && \text{not in the same direction.}
\end{aligned}
$$

The covariance between residuals and REs are given by:

$$
\begin{aligned}
Cov(r^*_{sdi}, \alpha_s) = Cov(\alpha^*_s + \alpha^*_{d|s} + \epsilon^*_{sdi}, \alpha_s) &= Cov(\alpha^*_s, \alpha_s) &&= \rho_{\alpha_s}\sigma_s\sigma^*_s &&= \sigma^*_{ss}, \\
Cov(r^*_{sdi}, \alpha_{d|s}) = Cov(\alpha^*_s + \alpha^*_{d|s} + \epsilon^*_{sdi}, \alpha_{d|s}) &= Cov(\alpha^*_{d|s}, \alpha_{d|s}) &&= \rho_{\alpha_{d|s}}\sigma_d\sigma^*_d &&= \sigma^*_{dd}.
\end{aligned}
$$

Therefore, the joint covariance of r_* and u is described by the following variance components:

$$
D = \begin{bmatrix} \sigma_s^2 & 0 & 0 \\ 0 & \sigma_d^2 & 0 \\ 0 & 0 & \sigma_d^2 \end{bmatrix},
$$

$$
W = \sigma^2_{s*} + \begin{bmatrix}
\sigma_d^{*2} + \sigma^{*2} & \sigma_d^{*2} & \sigma_d^{*2} & 0 & 0 & 0 \\
\sigma_d^{*2} & \sigma_d^{*2} + \sigma^{*2} & \sigma_d^{*2} & 0 & 0 & 0 \\
\sigma_d^{*2} & \sigma_d^{*2} & \sigma_d^{*2} + \sigma^{*2} & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_d^{*2} + \sigma^{*2} & \sigma_d^{*2} & \sigma_d^{*2} \\
0 & 0 & 0 & \sigma_d^{*2} & \sigma_d^{*2} + \sigma^{*2} & \sigma_d^2 \\
0 & 0 & 0 & \sigma_d^{*2} & \sigma_d^{*2} & \sigma_d^{*2} + \sigma^{*2}
\end{bmatrix},
$$

$$
C = \begin{bmatrix}
\sigma^*_{ss} & \sigma^*_{ss} & \sigma^*_{ss} & \sigma^*_{ss} & \sigma^*_{ss} & \sigma^*_{ss} \\
\sigma^*_{dd} & \sigma^*_{dd} & \sigma^*_{dd} & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma^*_{dd} & \sigma^*_{dd} & \sigma^*_{dd}
\end{bmatrix}.
$$

Note that D is the same as before, but W is different because the random effects in r are from the simpler model, and C is also affected.

Tables 2 through 9 report predicted speed deciles (pred.) and compare them with the observed ones (obs.).

Table 2: Predicted quantiles sections 1,3,5 and 7

	Section 1				Section 3				Section 5				Section 7			
quantile	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	50	56	48.6	54	65.2	58	70.4	63	59.3	60	53.5	57	65.4	66	66.1	72.6
20	56.7	59	56.2	58	68.7	63	73.1	68	64.7	64	61	63	70.7	69	71.8	76
30	61.5	62	61.6	60	71.2	66	75	71	68.6	68	66.3	67	74.5	73	76	79.8
40	65.6	65.8	66.3	62	73.3	69	76.7	74	72	71	70.9	70	77.8	77	79.5	82
50	69.5	67	70.7	65	75.3	72	78.2	77	75.1	73	75.2	75	80.9	80	82.8	85
60	73.3	69	75	67	77.3	74	79.8	79	78.3	77	79.5	77	84	82	86.1	87
70	77.4	71	79.7	70	79.4	77	81.5	82	81.6	81	84.1	81.9	87.3	86.8	89.6	89
80	82.3	73	85.2	74	81.9	81	83.4	85	85.5	85	89.5	87	91.1	90	93.8	93
90	88.9	77.3	92.8	79	85.3	88	86.1	89	91	90	97	94	96.5	95	99.5	97.4

Table 3: Predicted quantiles sections 8,9,10,11

quantile	Section 8				Section 9				Section 10				Section 11			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	68.2	69.4	67.9	74.2	63.2	60	65.2	70	47.4	50	54.5	56	56.4	59.3	56.4	59.3
20	73.9	75.4	73.5	77	68.4	67	70.4	72.8	51.8	51.2	57.2	60	62.5	63.2	62.5	61.7
30	78	81	77.5	77.6	72.1	68	74.1	76	54.9	54.6	59.1	63	66.9	65.1	66.8	67.1
40	81.5	84	81	81.6	75.2	73	77.3	79.6	57.6	55	60.8	64	70.6	68.1	70.6	71
50	84.8	86	84.2	85	78.2	75	80.3	82	60.1	56	62.4	65	74.2	70.6	74	72.5
60	88.1	88.2	87.4	87.2	81.1	79	83.3	86.4	62.6	57	63.9	67	77.7	71.9	77.5	74
70	91.6	90	90.9	90	84.3	86	86.5	88	65.3	59	65.6	68	81.5	75.9	81.2	76.4
80	95.7	94.2	94.9	94.8	87.9	87	90.2	91.2	68.4	60.8	67.5	71.2	85.9	81.6	85.6	81.6
90	101.4	103.5	100.5	100.6	93.1	90	95.4	94.1	72.7	64.4	70.2	73	92	95.6	91.6	93

Table 4: Predicted quantiles sections 12,13,14,15

quantile	Section 12				Section 13				Section 14				Section 15			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	60.9	59.7	61.1	59	62.9	62.7	62.2	65.8	64.5	67.1	63.4	67.8	64.6	64.3	63.7	65.7
20	66.9	64.2	67.4	67.5	68.6	67.4	68.3	68.5	71.2	70.6	70.6	72.1	70.7	69.5	70.2	73.8
30	71.3	71.3	71.8	72	72.7	70.7	72.7	72	76	76.3	75.7	75.3	75	74.3	74.8	75.6
40	75	73.9	75.7	78.3	76.3	75.4	76.5	74.3	80.1	79.6	80.2	79.1	78.8	78	78.7	80.2
50	78.5	76.9	79.3	82.9	79.6	79.8	80	81.9	83.9	83.1	84.3	83	82.2	81.9	82.4	83.8
60	81.9	85.7	82.8	86	82.9	84.2	83.5	84.4	87.7	85.9	88.5	90.6	85.7	84.3	86.1	85.2
70	85.7	88.8	86.7	90	86.4	87.8	87.3	84.8	91.8	91	92.9	94.1	89.4	88.6	90.1	88.7
80	90	92.1	91.2	91	90.5	89.1	91.7	86.9	96.6	96.3	98.1	97.9	93.8	92.6	94.7	93.7
90	96.1	94.8	97.4	106.8	96.3	93.2	97.8	93.4	103.3	107.1	105.3	105.9	99.8	103.1	101.1	97.2

Table 5: Predicted quantiles sections 16,17,18,21

quantile	Section 16				Section 17				Section 18				Section 21			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	65	71.7	64.4	71.3	62.4	66.2	64.2	67.8	51.4	51.9	51.2	52.6	85.8	73	86.2	81.7
20	71.8	75.2	71.1	74.6	69.1	69.2	70.3	70.9	57.8	54.2	57.8	56.6	89.3	88	90.3	89.2
30	76.6	80.2	76	76.3	74	72.9	74.8	74.4	62.4	58.4	62.6	60.9	91.9	92.5	93.2	95.3
40	80.8	82.5	80.2	79.6	78.1	74.5	78.6	77.4	66.3	65.3	66.7	62.2	94	97	95.6	98.4
50	84.7	85	84	82.2	82	76.9	82.2	80	70	67.2	70.5	68.5	96.1	98	98	104
60	88.6	88.3	87.9	84.4	85.9	85.4	85.7	83.3	73.7	69.7	74.3	71.8	98.1	99	100.3	105
70	92.7	90.5	92.1	91.5	90.1	92.3	89.6	88.7	77.6	74.9	78.4	75.4	100.3	99.5	102.8	105
80	97.6	93.5	96.9	94.3	95	96.3	94	92.4	82.2	80.3	83.1	84.9	102.8	100	105.7	111.8
90	104.3	108.6	103.7	100.2	101.7	99.3	100.2	94.1	88.6	91.2	89.7	91.7	106.3	105	109.7	114.9

Table 6: Predicted quantiles sections 22,23,24,25

quantile	Section 22				Section 23				Section 24				Section 25			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	60.8	64.8	56.3	64	45.5	39.1	43.7	46	63.6	67	66.6	59.7	53.7	54	55.5	50.4
20	68	69	64	70	48.3	42.4	47.6	47	69.1	68.6	71.2	66	58.7	56.2	60.5	55.2
30	73.2	71.4	69.6	71	50.3	46.1	50.5	48	73.1	69	74.6	75.3	62.2	59.8	64.1	65.4
40	77.6	78	74.4	72	52	46.8	52.9	50	76.4	72.4	77.5	77.4	65.3	62.6	67.2	71.4
50	81.8	82	78.9	72	53.6	47.5	55.2	50.5	79.6	74	80.1	79	68.1	66	70.1	73
60	86	94	83.3	75	55.2	48.2	57.4	51.8	82.7	79.4	82.8	81.8	71	68.6	73	78.4
70	90.4	97.8	88.1	76	57	48.9	59.8	54.1	86.1	81	85.7	88.1	74	71	76	81.2
80	95.6	101	93.7	79	59	50.2	62.7	57.4	90	86	89	90.2	77.6	75	79.7	84.2
90	102.8	108.4	101.5	82	61.7	51	66.6	61.5	95.5	99	93.7	96.1	82.5	75	84.7	89

Table 7: Predicted quantiles sections 26,27,28,29

quantile	Section 26				Section 27				Section 28				Section 29			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	57.8	59.4	60.8	57.2	34.8	44.6	35.4	45.8	66.5	67	68.3	76.6	68.2	72.3	67.2	78.5
20	63.5	65.6	65.4	60.6	40.4	46.2	41	47.6	73.3	75	74.7	78	75.2	77.1	74.8	80.1
30	67.5	67	68.8	65.5	44.4	47.8	45	48.4	78.1	77	79.4	83.8	80.2	84	80.3	82.6
40	71	67.6	71.6	68.8	47.9	49.4	48.4	49.2	82.3	78	83.4	86.4	84.5	89.2	85	85.1
50	74.2	73	74.3	76	51.1	51	51.6	51	86.2	87	87.1	87	88.5	93.1	89.4	85.1
60	77.5	75.8	76.9	76.8	54.3	53.6	54.8	53.6	90	87	90.8	87.6	92.6	93.1	93.8	93.7
70	81	77.1	79.8	77.3	57.7	54.2	58.2	55.6	94.2	88.5	94.8	91	96.9	93.1	98.5	101.2
80	85	80.8	83.1	80.4	61.7	55	62.2	56.8	99	89	99.4	96.8	101.9	94.9	103.9	103.9
90	90.7	86.7	87.7	86.6	67.3	60	67.8	60.2	105.8	103	105.9	100.8	108.9	109.7	111.5	108.8

Table 8: Predicted quantiles sections 30,31,32,33

quantile	Section 30				Section 31				Section 32				Section 33			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	62.6	56.3	65.3	57.4	56.8	61.2	60.6	67.2	46.2	52.1	47.4	55.2	54	52.7	56.5	58.2
20	64.2	56.3	66.5	62.1	62.9	66.5	65.6	68.9	52.9	57.5	53.7	59.2	58.8	58.1	60.6	60.9
30	65.3	57.5	67.4	62.1	67.2	68	69.3	70	57.7	59.6	58.3	62.2	62.2	60	63.6	64
40	66.3	61.4	68.1	63.8	70.9	70.3	72.3	73	61.9	59.6	62.2	63.9	65.2	62.7	66.1	67.4
50	67.3	65.9	68.8	71.8	74.4	71.9	75.2	74.4	65.7	64.2	65.8	67.4	67.9	66.2	68.5	67.4
60	68.2	69.3	69.5	74.1	77.9	73.2	78.1	77.1	69.6	66.7	69.4	69.3	70.7	67.6	70.9	67.4
70	69.2	69.3	70.2	74.1	81.6	74.6	81.2	79.6	73.7	69.5	73.3	71.4	73.6	71.1	73.4	71.1
80	70.3	87.2	71.1	79.2	85.9	76.9	84.8	83.4	78.5	79.4	77.9	73.5	77.1	73.8	76.4	71.9
90	71.9	93.2	72.3	79.2	92	84.1	89.8	87.5	85.2	83.4	84.2	77.3	81.9	81.3	80.5	75.3

Table 9: Predicted quantiles sections 34,36

quantile	Section 34				Section 36			
	dir 1		dir 2		dir 1		dir 2	
	pred.	obs.	pred.	obs.	pred.	obs.	pred.	obs.
10	45.6	42.5	43.9	43.1	45.3	44.6	46.5	56.7
20	49.3	46.5	48.7	50.6	53.1	50.9	53.8	59.1
30	52	49.1	52.2	52.7	58.8	59.2	59.1	62.2
40	54.3	51.2	55.2	56.2	63.6	64.2	63.6	65.2
50	56.5	53.2	57.9	58.8	68.1	68.2	67.8	69.5
60	58.6	54.7	60.7	60.2	72.6	70.9	72	72.4
70	60.9	58	63.7	64.9	77.5	74.9	76.5	76.8
80	63.6	62.2	67.2	70.3	83.1	79.3	81.8	78.8
90	67.3	68.4	72	76	91	94.6	89.1	82.5

When analyzing Tables 2 to 9, a number of patterns can be observed. First, the speed quantiles of most sections are predicted accurately, which attests the validity of our statistical methodology. For example, most of the sections have predictions that are within an interval of 2-3 kilometer per hour (km/h) and for almost all the quantiles. Most central deciles (40^{th} to 60^{th} quantiles) have very good predictions and therefore it is possible to say that additional observations collected to predict the REs can be successfully used to predict the mean speed of the section with some accuracy. The worst value obtained for the predicted median (50^{th} quantile) is the one related to section 29, direction 1, with an observed median of 93.1 km/h and a prediction of 88.5 km/h. This compares with the 10^{th} quantile that predicts 68.2 km/h for an observed value of 72.3 km/h.

Second, some sections are likely to have an error in part due to the prediction error of the random effects. One way to assess this is to observe the extreme quantiles of those sections for which both the 10^{th} and the 90^{th} quantiles are underestimated. Example of this type is section 22, direction 1. Looking back at the plots of predicted REs in figure 3, it is possible to observe that the direction effects are not very precise with only five observations and this is likely a case where the prediction has created a small error. For section 23 (figure 4) the section effect and one of the direction effect are overestimated with five observations, which would explain this component of the prediction error.

Third, the most obvious prediction error is the overestimation of low deciles and the underestimation of high deciles, or vice-versa. This can be observed for example in section 10, direction 1 for which the 10^{th} quantile is underestimated by 2.6 km/h while the 90^{th} quantile is overestimated by 8.3 km/h. There might still be an error in the predicted RE for this section but it cannot be fixed because the RE is a constant added to all predictions in the same section and direction. The most likely cause for this kind of error lies in the estimated coefficients. This possible source of error will be investigated in the next section using a re-sampling technique that will determine if data from a particular section has a relevant effects on the values of the coefficients estimated.

5 Jackknife coefficients

A jackknife re-sampling technique (Efron and Tibshirani 1993) is adopted to explore the poor predictions obtained for some combinations of sections and directions; we are particularly interested in predictions of high and low speed quantiles. We have already discussed in the previous section how to compute out of sample predictors for REs. However, the prediction of speed quantiles in a new section also requires the model's coefficients to be estimated for the validation sample.

The use of a jackknife procedure is based on the following observations: (1) the modeling approach suggested in this paper involves the fitting of a model on a set of road sections and the application of it to another section for which data is not available; (2) some sections are poorly predicted by the suggested model but (3) the majority of sections are predicted with remarkable accuracy.

We are specifically interested in comparing the jackknife coefficients obtained by excluding a generic section s and the quality of the forecasts for that section. We hypothesize that significant discrepancies among the values predicted and observed indicate the existence of one or more of the following problems: (i) inaccuracies in the data, (ii) errors in data collection (perhaps these sections have been sampled by operators who did not get proper training, and/or did not assess accurately if the vehicle was traveling under free flow conditions), (iii) different driving conditions might alter motorists' behavior, (iv) the omission of important covariates that were overlooked and that affect some sections. These scenarios would generally cause some sections to be poorly modeled by the approach suggested in this paper and the assessment of such conditions is expected to greatly simplify the investigation task as the model's user validates his data.

In a model that is appropriate for its data, it is expected that the exclusion of any observation does not change the estimates of the model coefficients in a relevant way. The re-sampling scheme suggested discards a substantial number of observations. This way to proceed makes it impossible to directly use the standard re-sampling literature to quantify the effect of a specific coefficient, and ultimately to identify the specification issue. Furthermore, the use of such a criteria would involve a different threshold for each section due to the variability of their size. Simulations could be used to derive a more systematic identification method for sections requiring attention, however this is outside the scope of our analysis and we rely instead on a qualitative evaluation of the effect of each section on coefficients' estimates.

Tables 10, and 11 in Appendix A present the jackknife coefficients used for speed data validation and should be read together. Each line corresponds to estimates obtained by removing one section, except the first one that contains the estimations on the whole sample. Each column provides the estimated coefficients for a specific predictor. The cells in bold indicate the coefficients that are very different from the full sample equivalent (outside two standard deviations). Ideally, the estimated coefficients in a given column would be stable and comparable to the coefficient estimated using the entire sample. Large differences between a coefficient estimated by excluding one section and those obtained with the whole sample raise concerns about the specific section or the predictor.

When reading the results, it can be observed that the jackknife coefficients for `Z*SRW` are very similar across the thirty one jackknife samples and are very similar to the coefficients estimated on the overall sample, whereas the coefficients for `Z*1/R` appear to be off only for sections 7 and 10. Sections 1, 3, 7, and 36 generate some of the extreme jackknife coefficients. We can probably conclude that these sections (1,3,7, and 36) behave according to a different model, or that the measurement of predictors and speed data was less reliable on these sections. This is also confirmed in the comparison between predition and observation reported in the Tables 2 to 9.

6 Conclusions

In this paper we have explored methods to predict speed distributions using mixed linear models. Difficulties associated with this problem are twofold. First we rely on model with random effects to make prediction on new road sections. Second, we are using normal quantiles of the dependent variables as predictors.

We have discussed that it is necessary to observe at least some speed data in the section for which decile predictions were computed. To overcome the problem created by the use of normal quantiles in the calculation of residuals, we propose the use of an auxiliary model. It has been shown how the relationship between this auxiliary model and the full actual model can be used in prediction in order to derive the *best linear unbiased predictor* (BLUP) in this context. It was also observed that this method was not performing well when used to make random effect predictions for a model that ignored the sampling design of the data, but turned out to be very precise when the full sampling design was accounted in the model.

The approach was further tested to validate the models' coefficients associated with the variables of the model. The estimates obtained were roughly stable except for some variables that generated more extreme coefficients for some sections. The sections with extreme coefficients were consistently the same for all the affected variables.

A jackknife technique was used to understand what road sections caused poor predictions of the random effects. It was found that most road sections were predicted satisfactorily. Large errors in the prediction of random effect were mostly caused by errors in the coefficients of the model. Some sections appeared to have a disproportionate effect on the model, suggesting that they should be modeled in a different way.

Overall, this paper has contributed to the transferability of RE models, has identified the problems arising in the estimation of Operating Speeds, has developed a theory to calculate the best predictors and to validate the results obtained. Future efforts should be directed towards the use of the method proposed in practice and possibly to different model types that include REs.

ACKNOWLEDGEMENTS

The research work included in this paper was made possible thanks to the joint project between the Politecnico di Torino (Italy) and the University of Maryland (US). The authors acknowledge funding received from the Compagnia di San Paolo (Italy) and from the Politecnico di Torino (Rectoral Decree n. 208 of the 24th of May, 2013). The Città Metropolitana di Torino, and the Provinces of Vercelli and Alessandria are greatly acknowledged for the access to the GIS databases used to collect roadway inventory information.

Appendices

Table 10: Jackknife coefficients - Part 1

section out	Intercept	PedD	1/R	Z*Lane	Z*1/R	Z*SLS	Z*SBLS	Z*IDRS	Z*SRS	Z*SBRS	Z*SRW	Z*SLW
none	79.33	-7.59	-1946.46	20.54	-274.07	-3.46	-1.56	-0.83	-2.88	-0.64	-13.38	13.85
1	79.8	-7.77	-2042.39	**25.07**	-227.05	**-1.95**	**-2.82**	-0.91	**-0.79**	**-1.36**	-13.62	**13.09**
3	79.51	-7.67	-1983.47	22.96	**-170.71**	**-2.49**	-1.19	**-0.97**	-2.03	-0.31	-12.56	**14.29**
5	79.48	-7.65	-1976.76	19.71	-279.35	-3.38	-2.24	-0.95	-3.1	**0.12**	-13.41	13.74
7	79.21	-7.55	**-2277.82**	21.89	-289.06	-3.45	**-0.93**	-0.83	-2.72	-0.47	-13.36	14.13
8	79.09	-7.5	-1896.77	20.86	-283.21	-3.44	-1.44	-0.83	-2.86	-0.57	-13.36	13.9
9	79.3	-7.58	-1943.43	20.46	-274.26	-3.48	-1.64	-0.84	-2.88	-0.59	-13.4	13.87
10	79.2	-7.54	**-1208.28**	21.33	-248.33	-3.31	-1.48	-0.81	-2.82	-0.54	-13.34	13.78
11	79.53	-7.67	-1987.44	19.97	-267.14	-3.45	-1.5	-0.8	-2.88	-0.71	-13.37	13.85
12	79.28	-7.57	-1936.44	18.9	-271.29	-3.47	-1.53	-0.82	-2.94	-0.66	-13.37	13.89
13	79.33	-7.59	-1946	21.35	-283.83	-3.46	-1.58	-0.84	-2.88	-0.58	-13.39	13.85
14	79.12	-7.51	-1901.82	20.27	-274.76	-3.45	-1.57	-0.83	-2.91	-0.61	-13.36	13.84
15	79.22	-7.55	-1922.98	20.74	-273.66	-3.45	-1.59	-0.86	-2.88	-0.61	-13.38	13.85
16	79.12	-7.51	-1901.59	21.76	-266.84	-3.43	-1.58	-0.87	-2.83	-0.64	-13.38	13.82
17	79.23	-7.55	-1924.69	20.26	-270.36	-3.46	-1.6	-0.86	-2.88	-0.62	-13.38	13.85
18	79.68	-7.72	-2017.04	21.32	-276.61	-3.41	-1.49	-0.82	-2.84	-0.56	-13.4	13.84
20	**78.42**	-7.24	-1757.97	19.7	-278.04	-3.52	-1.65	-0.83	-2.93	-0.72	-13.37	13.87
21	78.71	-7.36	-1819.11	21.88	-244.19	-3.32	-1.37	-0.9	-2.78	-0.51	-13.39	13.82
22	79.33	-7.59	-1944.68	20.45	-274.83	-3.46	-1.5	-0.81	-2.94	-0.7	-13.37	13.89
23	**80.41**	-8	-2168.41	20.52	-276.31	-3.46	-1.55	-0.83	-2.88	-0.64	-13.38	13.86
24	79.33	-7.59	-1945.62	20.49	-273.89	-3.55	-1.54	-0.82	-2.86	-0.66	-13.4	13.84
25	79.73	-7.74	-2028.57	20.52	-273.66	-3.47	-1.57	-0.83	-2.89	-0.66	-13.39	13.85
26	79.58	-7.69	-1997.16	20.52	-273.93	-3.59	-1.46	-0.79	-2.74	-0.71	-13.38	13.88
27	79.48	**-8.74**	-1976.04	20.47	-269.15	-3.5	-1.59	-0.86	-2.83	-0.59	-13.35	13.85
28	79.09	-7.5	-1895.65	19.77	-272.6	-3.5	-1.67	-0.85	-2.91	-0.7	-13.36	13.85
29	78.93	-7.44	-1863.92	20.07	-267.62	-3.49	-1.65	-0.85	-2.93	-0.73	-13.34	13.86
30	79.37	-7.56	-1954.05	19.1	-214.59	**-4.27**	-2.05	**-0.99**	-3.53	-1.15	-13.43	13.63
31	79.52	-7.66	-1983.91	20.34	-256.36	-3.1	-1.63	-0.87	-2.74	-0.71	-13.26	13.93
32	79.47	-7.43	-1973.31	21.44	-295.67	-3.71	-1.6	**-0.7**	-3.11	-0.74	-13.36	13.82
33	79.78	-7.76	-2037.72	21.32	-276.67	-3.33	-1.5	-0.81	-3.05	-0.6	-13.39	13.83
34	79.34	-7.66	-1947.25	19.84	-275.81	-3.59	-1.67	-0.79	-3.01	-0.76	-13.35	13.87
36	79.39	-7.52	-1958.78	**17.03**	-257.65	**-4.26**	**-2.33**	-0.91	-3.62	-1.11	-13.31	13.82

Table 11: Jackknife coefficients - Part 2

section out	Z*PLS	Z*DDRS	Z*TRDLS	Z*LW	Z*IDLS	Z*LG	Z*ΔPSL	Z*PedD	Z*TRDRS	Z*LBRS	Z*LBLS
none	0.06	0.43	-0.29	-3.52	-0.43	0.07	0.07	1.62	0.41	1.31	1.16
1	0.04	**0.26**	**-0.78**	-4.16	**-0.3**	0.09	**0.19**	**0.11**	0.03	**2.57**	**2.52**
3	**0.16**	**0.6**	**-0.82**	**-6.73**	-0.52	0.06	0.03	2.13	**-0.3**	1.03	1.25
5	0.07	0.45	-0.27	-3.37	-0.4	0.11	0.06	1.71	0.34	1.53	0.87
7	0.07	0.44	**-0.86**	-4.13	-0.46	**-0.08**	0.05	1.66	0.04	1.67	1.09
8	0.06	0.43	-0.4	-3.67	-0.43	0.06	0.07	1.62	0.31	1.36	1.17
9	0.06	0.42	-0.31	-3.48	-0.42	0.05	0.07	1.62	0.42	1.33	1.13
10	0.06	0.44	-0.25	-3.77	-0.49	0.13	0.07	1.58	0.5	1.19	1.17
11	0.06	0.43	-0.27	-3.37	-0.49	0.08	0.07	1.63	0.4	1.32	1.12
12	0.06	0.47	-0.24	-3.17	-0.42	0.07	0.08	1.61	0.4	1.35	1.26
13	0.06	0.42	-0.3	-3.75	-0.39	0.07	0.07	1.59	0.4	1.3	1.18
14	0.06	0.44	-0.26	-3.48	-0.42	0.09	0.07	1.6	0.39	1.3	1.22
15	0.06	0.43	-0.29	-3.58	-0.41	0.07	0.07	1.62	0.41	1.29	1.18
16	0.06	0.39	-0.32	-3.78	-0.46	0.07	0.07	1.65	0.41	1.26	1.06
17	0.06	0.42	-0.28	-3.44	-0.42	0.07	0.07	1.63	0.41	1.29	1.17
18	0.06	0.44	-0.3	-3.83	-0.42	0.07	0.08	1.59	0.37	1.39	1.27
20	0.06	0.4	-0.28	-3.18	-0.43	0.07	0.07	1.63	0.44	1.24	1.03
21	0.06	0.46	-0.46	-4	-0.51	0.07	0.07	1.67	0.17	1.38	1.3
22	0.06	0.45	-0.27	-3.53	-0.45	0.06	0.07	1.6	0.41	1.41	1.08
23	0.06	0.42	-0.3	-3.5	-0.43	0.07	0.07	1.61	0.39	1.31	1.14
24	0.06	0.41	-0.3	-3.5	-0.45	0.07	0.07	1.67	0.41	1.34	1.11
25	0.06	0.43	-0.28	-3.52	-0.43	0.06	0.07	1.62	0.42	1.33	1.16
26	0.06	0.44	-0.29	-3.53	-0.48	0.06	0.07	1.59	0.39	1.38	1.08
27	0.06	0.36	-0.33	-3.5	-0.44	0.07	0.06	2.12	0.46	1.33	1.03
28	0.06	0.4	-0.28	-3.19	-0.44	0.08	0.07	1.65	0.43	1.23	1.03
29	0.06	0.43	-0.12	-3.43	-0.45	0.07	0.08	1.65	0.57	1.28	1.14
30	0.05	0.34	-0.37	-2.7	-0.53	0.06	0.1	1.99	0.43	1.44	1.18
31	0.06	0.46	-0.27	-3.39	-0.46	0.07	0.08	1.38	0.38	1.21	1.12
32	0.06	0.38	-0.33	-3.75	-0.41	0.08	0.08	1.76	0.38	1.6	1.13
33	0.06	0.43	-0.31	-3.81	-0.43	0.07	0.08	1.63	0.37	1.41	1.24
34	0.06	0.41	-0.3	-3.19	-0.44	0.07	0.08	1.25	0.39	1.28	1.1
36	0.04	0.43	-0.35	**-1.76**	-0.38	0.1	0.1	1.81	0.28	0.82	1.29

References

Bassani, M., L. Catani, C. Cirillo, and G. Mutani (2016). "Night-time and daytime operating speed distribution in urban arterials." In: *Transportation Research Part F: Traffic Psychology and Behaviour* 42.1, pp. 56–69.

Bassani, M., C. Cirillo, S. Molinari, and J.M. Tremblay (2016). "Random effect models to predict operating speed distribution on rural two-lane highways." In: *Journal of Transportation Engineering* 142.6, p. 04016019.

Bassani, M., D. Dalmazzo, G. Marinelli, and C. Cirillo (2014). "The effects of road geometrics and traffic regulations on driver-preferred speeds in northern Italy. An exploratory analysis." In: *Transportation Research Part F: Traffic Psychology and Behaviour* 25, pp. 10–26.

Catani, L., J.M. Tremblay, M. Bassani, and C. Cirillo (2017). "Methodology to backcalculate individual speed data originally aggregated by road detectors." In: *Transportation Research Record: Journal of the Transportation Research Board* 2659.1, pp. 1–14.

Cheng, W., G.S. Gill, T. Sakrani, D. Ralls, and X. Jia (2018). "Modeling the endogeneity of lane-mean speeds and lane-speed deviations using a bayesian structural equations approach with spatial correlation." In: *Transportation Research Part A: Policy and Practice* 116, pp. 220–231.

Dimaiuta, M., E. Donnell, S.C. Himes, and R. J. Porter (2011). "Modeling operating speed." In: *Transportation Research E-circular* E-C151.

Efron, B. and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman et Hall.

Figueroa-Medina, A. and A.P. Tarko (2005). "Speed factors on two-lane rural highways in free-flow conditions." In: *Transportation Research Record: Journal of the Transportation Research Board* 1912, pp. 39–46.

Garber, N. and L. Hoel (2020). *Traffic and highway engineering.* 5th Edition. Cengage Learning.

Himes, S., E. Donnell, and R. Porter (2013). "Posted speed limit: To include or not to include in operating speed models." In: *Transportation Research Part A: Policy and Practice* 52, pp. 23–33.

Islam, M.T. and K. El-Basyouny (2015). "Multilevel models to analyze before and after speed data." In: *Analytic Methods in Accident Research* 8, pp. 33–44.

Lamm, R., E.M. Choueiri, J.C. Hayward, and Paluri. A. (1988). "Possible design procedure to promote design consistency in highway geometric design on two-lane rural road." In: *Transportation Research Record: Journal of the Transportation Research Board* 1195, pp. 111–122.

McCulloch, C.E., S.R. Searle, and J.M. Neuhaus (2008). *Generalized, Linear, and Mixed Models*. Second. Hoboken, New-Jersey: John Wiley & Sons.

McFadden, J. and L. Elefteriadou (2000). "Evaluating horizontal alignment design consistency of two lane rural highways: development of new procedure." In: *Transportation Research Record: Journal of the Transportation Research Board* 1737, pp. 9–17.

Park, Y.J. and F.F. Saccomanno (2006). "Evaluating speed consistency between successive elements of a two-lane rural highway." In: *Transportation Research Part A: Policy and Practice* 40.5, pp. 375–385.

Shankar, V. and F. Mannering (1998). "Modeling the endogeneity of lane-mean speeds and lane-speed deviations: a structural equations approach." In: *Transportation Research Part A: Policy and Practice* 32.5, pp. 311–322.

Tarris, J., C. Poe, J. Mason, and K. Goulias (1996). "Predicting operating speeds on low-speed urban streets: Regression and panel analysis approaches." In: *Transportation Research Record: Journal of the Transportation Research Board* 1523, pp. 46–54.