Prediction of regionalized car insurance risks based on control variates

Marcus C. Christiansen, Christian Hirsch und Volker Schmidt

Preprint Series: 2013 - 03

Fakultät für Mathematik und Wirtschaftswissenschaften
UNIVERSITÄT ULM

Prediction of regionalized car insurance risks based on control variates

Marcus C. Christiansen, Christian Hirsch, Volker Schmidt

March 26, 2013

Abstract

We show how regional prediction of car insurance risks can be improved by combining explanatory modeling with phenomenological models from industrial practice. Motivated by the control-variates technique, we propose a suitable combined predictor. We provide explicit conditions which imply that the mean squared error of the combined predictor is smaller than the mean squared error of the standard predictor currently used in industry and smaller than predictors from explanatory modeling. We also discuss how a non-parametric random forest approach may be used to practically compute such predictors and consider an application to German car insurance data.

1 Introduction

In Germany, car insurance premiums heavily depend on regional classifications. In practice, premiums are multiplied by factors subject to the residence of the car owner. These multipliers are estimated on the basis of statistical observations from the recent past. Few car insurers, if any, have the necessary data to estimate multipliers all over Germany. Therefore, the German Insurance Association (GDV) pools claims data from its members in a central database and creates regional classifications that are available for all insurers. Although the GDV classification divides Germany into more than 400 regions, there appear erratic transitions at regional borders. For example, we observe unrealistic differences in the risk multipliers of the eastern outskirts of Berlin and their neighboring regions such as the Oder-Spree county. Practitioners intuition suggests that these neighboring districts, which are very similar with regard to urban development and socio-economic circumstances, should have similar multipliers.

While the GDV uses a purely phenomenological modeling, better results might be expected from explanatory models. However, as far as we know,

AMS 1991 subject classification: Primary: 62P05; Secondary: 91G99
Key words and phrases: prediction, regionalized risk, car insurance, random forest, control variates

there exist no explanatory models that come close enough to the real data (nevertheless, see [6] for a more refined statistical approach to predict regional risk levels phenomenologically). In the present paper we propose a method that combines an explanatory approach with the phenomenological GDV model. The idea is to make use of an explanatory component as far as possible and to supplement it with the GDV model in order to incorporate effects that can not be explained so far. Our approach is closely related to the so-called *control-variate technique* which constitutes a popular means of variance reduction in Monte-Carlo simulations, see e.g. [5, 8]. Within our modeling framework we can prove that our mixed approach is in some sense optimal.

In our application to German car insurance data presented in Section 5 we solely focus on third party insurance, which is compulsory in Germany since 1939. Indeed, third party insurance data is particularly well-suited for our methods, since significant correlations between regional risks and publicly available road data can be observed. Yet, our concept can also be applied in comprehensive car insurance and might be helpful in other lines of business far beyond car insurances.

2 The regional classification of the GDV

In the present section we provide a brief description of the current approach, which is used in German insurance industry for the prediction of regional risk levels. As has already been explained in Section 1, in Germany it is common practice that the premium for car insurance (third party as well as comprehensive) depends on the county (so-called *Landkreis*) where the corresponding vehicle is registered. In order to obtain a reliable data-basis the GDV collects claim data of various insurers and associates with each county a *risk level*, which reflects the deviation of the claim sizes in the respective county from the federal average, see [4]. A risk level of 100 corresponds precisely to the federal average, while values below or above 100 correspond to more favorable or less favorable risk situations, respectively. The precise formula which is used to determine the risk level given the claim data is not publicly available, but it also incorporates averaging over the most recent years. The risk levels are used to define 12 risk classes. Figure 2.1 shows a map of German counties colored according to their risk class in third party car insurance.

3 Statistical model

3.1 Predictors of risk levels

Assume that we have a country that consists of n regions, where the i-th region has m_i subregions, for each $i = 1, \dots, n$. In our application, the regions $i = 1, ..., n$ will equal the GDV classification regions, which we refine into further subregions in order to make the regional classification more smooth. Let Θ_i and Θ_{ij} describe fundamental risk levels of region i and subregion ij, for each

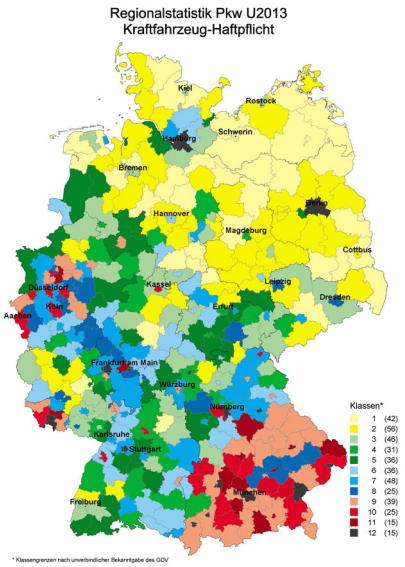

Figure 2.1: Regional risk classes in third party car insurance (source: GDV, [4])

$i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We take a Bayesian perspective here and assume that the risk levels are identically distributed random variables with finite second moment. For example, think of Θ_i and Θ_{ij} as claim costs per policy averaged over all policy holders in region i and in subregion ij, respectively. These values may change from one year to another due to certain random effects. Further, for some integer $d \geq 1$ let Δ_i and Δ_{ij} be d-dimensional explanatory covariates for region i and subregion ij. For example, Δ_i and Δ_{ij} could contain information on the road density. Taking these covariates into our Bayesian framework, we assume that (Θ_i, Δ_i), $i = 1, ..., n$ are identically distributed random vectors. Similarly we also assume that $(\Theta_{ij}, \Delta_{ij})$, $i = 1, ..., n$, $j = 1, ..., m_i$, are identically distributed random vectors. Suppose that covariate data are available for each region and subregion but that claim data are only available for the regions and not for the subregions. In our application, each insurer has access to the GDV classification for regions $i = 1, ..., n$ but does not have more detailed information for any subregions. That means that we can observe the Θ_i, Δ_i, and Δ_{ij} but not the Θ_{ij} . In insurance practice, Θ_{ij} is then typically predicted by

$$\Theta_{ij}^{(1)} = \Theta_i, \tag{3.1}$$

i.e., the premium of a car insurance of an owner that is registered in subregion ij is calculated with the general risk factor for the entire mother region i. Given the available information $\Theta_i = \theta_i$, $\Delta_i = \delta_i$, $\Delta_{ij} = \delta_{ij}$, the mean squared prediction

error is given by

$$E^{(\theta_i,\delta_i,\delta_{ij})}[(\Theta_{ij}-\Theta_{ij}^{(1)})^2]=E[(\Theta_{ij}-\Theta_i)^2|\Theta_i=\theta_i,\Delta_i=\delta_i,\Delta_{ij}=\delta_{ij}]. \quad (3.2)$$

In this context we investigate the following question: Can we find a predictor for Θ_{ij} that is better than Θ_i, i.e., a predictor whose mean squared prediction error is smaller than the value given in (3.2)? An alternative to the phenomenological model of the GDV is explanatory modeling. A natural idea is here to predict Θ_{ij} by $\Theta_{ij}^{(2)}=E[\Theta_{ij}|\Delta_{ij}]=f_{ij}(\Delta_{ij})$, since among all Δ_{ij}-measurable random variables Θ with finite second moment the mean squared error $E[|\Theta-\Theta_{ij}|^2]$ is minimized for $\Theta=\Theta_{ij}^{(2)}$.

The function $f_{ij}:\mathbb{R}^d\to\mathbb{R}$ describes the effect of the explanatory covariate Δ_{ij} on the regional risk level Θ_{ij}. Since we assumed that the (Θ_i,Δ_i), $i=1,\dots,n$ are identically distributed and also that the $(\Theta_{ij},\Delta_{ij})$, $i=1,...,n$, $j=1,...,m_i$, are identically distributed, there exist measurable functions $f,f':\mathbb{R}^d\to\mathbb{R}$ such that

$$\begin{aligned} E[\Theta_i|\Delta_i]&=f(\Delta_i),\\ E[\Theta_{ij}|\Delta_{ij}]&=f'(\Delta_{ij}) \end{aligned} \quad (3.3)$$

for all $i=1,...,n$ and $j=1,...,m_i$. We additionally assume that the functions f and f' coincide. In other words, the effects that the covariates Δ_{ij} and Δ_i have on Θ_{ij} and Θ_i, respectively, are the same over the whole country and do not vary by region. In our application this means for example that if a higher road density increases the risk in some region by some amount, then this is also true for all other regions. This seems to be a plausible assumption for car insurance. With using the same function f for both regions and subregions, we can estimate f solely from data on the regional level and then make use of that information on the subregional level. However, finding the function f is still not a trivial task. Section 4 explains how f can be estimated from $(\Theta_1,\Delta_1),...,(\Theta_n,\Delta_n)$. We finally obtain

$$\Theta_{ij}^{(2)}=f(\Delta_{ij}) \quad (3.4)$$

as a predictor for Θ_{ij}. This method works well in practice if $f(\Delta_{ij})$ explains most of the variance of Θ_{ij}. In our application to car insurance, we were able to explain about two-thirds of the variance, which is quite encouraging but not enough to clearly trump the phenomenological GDV model. We therefore discuss the question whether there is still another alternative.

3.2 Combined predictor

The GDV predictor $\Theta_{ij}^{(1)}=\Theta_i$ is motivated by the empirical observation that the subregional risk level Θ_{ij} positively correlates with the regional risk level Θ_i. Similarly, it is plausible to assume that the error $\Gamma_{ij}=\Theta_{ij}-f(\Delta_{ij})$ of

the predictor $\Theta_{ij}^{(2)}$ positively correlates with the (observable) quantity $\Gamma_i = \Theta_i - f(\Delta_i)$. Therefore we suggest to replace the predictor $\Theta_{ij}^{(2)}$ by

$$\Theta_{ij}^{(3)} = f(\Delta_{ij}) + \Theta_i - f(\Delta_i). \tag{3.5}$$

Our suggestion is motivated by the notion of control variates, a popular variance reduction method used in Monte-Carlo simulations, see e.g. [5, 8]. Originally, this approach is used to enhance the precision when computing the expectation $E[X]$ for some random variable X. Instead of considering the estimator X for $E[X]$ it is often convenient to consider estimators of the form $X + (E[Y] - Y)$ for some random variable Y, whose mean is analytically computable. If X and Y are highly correlated, then so are the errors $E[X] - X$ and $E[Y] - Y$ and it is plausible that $X + (E[Y] - Y)$ yields a better estimate for $E[X]$. We refer the reader to [5, 8] for further details on control variates.

We now derive conditions under which $\Theta_{ij}^{(3)}$ is a better predictor for Θ_{ij} than $\Theta_{ij}^{(1)}$ and $\Theta_{ij}^{(2)}$.

Proposition 3.1. *Assume that*

$$E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}] = E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_i]. \tag{3.6}$$

Then

$$\begin{aligned}
E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - \Theta_i)^2\big] &= (f(\delta_{ij}) - f(\delta_i))^2 + E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Gamma_{ij} - \Gamma_i)^2\big],\\
E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}) - \Gamma_i)^2\big] &= E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Gamma_{ij} - \Gamma_i)^2\big],\\
E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}))^2\big] &= E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}^2],
\end{aligned}$$

and for any $\lambda \in \mathbb{R}$

$$E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}) - \lambda\Gamma_i)^2\big] = E^{(\theta_i,\delta_i,\delta_{ij})}\big[\Gamma_{ij}^2\big] - (2\lambda - \lambda^2)\,(\theta_i - f(\delta_i))^2.$$

Proof. Note that we can rewrite $E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_i - \Theta_{ij})^2\big]$ as $E^{(\theta_i,\delta_i,\delta_{ij})}\big[(f(\delta_{ij}) + \Gamma_{ij} - f(\delta_i) - \Gamma_i)^2\big]$. Similarly, we can rewrite all other conditional expectations that occur in the proposition and in the remaining part of the proof. By expanding the expression $(f(\delta_{ij}) + \Gamma_{ij} - f(\delta_i) - \Gamma_i)^2$ and by using the fact that

$$E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}] = E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_i] = \theta_i - f(\delta_i),$$

we get that

$$\begin{aligned}
&E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - \Theta_i)^2\big] - (f(\delta_{ij}) - f(\delta_i))^2 - E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Gamma_{ij} - \Gamma_i)^2\big]\\
&= 2\,E^{(\theta_i,\delta_i,\delta_{ij})}\big[(f(\Delta_{ij}) - f(\Delta_i))(\Gamma_{ij} - \Gamma_i)\big]\\
&= 2\,(f(\delta_{ij}) - f(\delta_i))\,E^{(\theta_i,\delta_i,\delta_{ij})}\big[\Gamma_{ij} - \Gamma_i\big]\\
&= 0.
\end{aligned}$$

Analogously, by expanding the expression $(\Theta_{ij} - f(\Delta_{ij}) - \lambda\Gamma_i)^2 = (\Gamma_{ij} - \lambda\Gamma_i)^2$ we get that

$$\begin{aligned}
&E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}) - \lambda\Gamma_i)^2\big] \\
&= E^{(\theta_i,\delta_i,\delta_{ij})}\big[\Gamma_{ij}^2 - 2\,\lambda\,\Gamma_i\Gamma_{ij} + \lambda^2\Gamma_i^2\big] \\
&= E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}^2] - 2\,\lambda\,E^{(\theta_i,\delta_i,\delta_{ij})}[(\theta_i - f(\delta_i))\,\Gamma_{ij}] + \lambda^2 E^{(\theta_i,\delta_i,\delta_{ij})}[(\theta_i - f(\delta_i))^2] \\
&= E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}^2] - 2\,\lambda\,(\theta_i - f(\delta_i))^2 + \lambda^2(\theta_i - f(\delta_i))^2.
\end{aligned}$$

This completes the proof. □

Proposition 3.1 shows that predicting Θ_{ij} by $\Theta_{ij}^{(3)} = f(\Delta_{ij}) + \Gamma_i$ instead of considering the predictor $\Theta_{ij}^{(1)} = \Theta_i$ always leads to a decrease in mean squared error. Moreover, the following result is true.

Corollary 3.2. *Under assumption* (3.6)*, it holds that*

$$E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - \Theta_{ij}^{(3)})^2\big] = \min_{\lambda\in\mathbb{R}} E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}) - \lambda\Gamma_i)^2\big]. \tag{3.7}$$

Furthermore,

$$E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - \Theta_{ij}^{(2)})^2\big] \le E^{(\theta_i,\delta_i,\delta_{ij})}\big[(\Theta_{ij} - \Theta_{ij}^{(1)})^2\big]$$

if and only if $(\theta_i - f(\delta_i))^2 \le (f(\delta_{ij}) - f(\delta_i))^2$.

Corollary 3.2 shows that $\Theta_{ij}^{(3)}$ is the best estimator among all linearly combined estimators of the form $f(\Delta_{ij}) + \lambda\Gamma_i = f(\Delta_{ij}) + \lambda(\Theta_i - f(\Delta_i))$, $\lambda \in \mathbb{R}$. In particular, $\Theta_{ij}^{(3)}$ is always better than $\Theta_{ij}^{(2)}$ provided that (3.6) holds. In the special case $E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}] = 0$, which is here equivalent to $\theta_i - f(\delta_i) = 0$, the mean squared error for the predictor $\Theta_{ij}^{(2)}$ of the pure explanatory model equals the mean squared error for $\Theta_{ij}^{(3)}$. Comparing the GDV predictor $\Theta_{ij}^{(1)}$ and the pure explanatory predictor $\Theta_{ij}^{(2)}$, there is no clear winner.

The assumption (3.6), i.e., $E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}] = E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_i] = \theta_i - f(\delta_i)$, implies that the mean level of Γ_{ij} is always $\theta_i - f(\delta_i)$, regardless of the covariate Δ_{ij} and regardless of j. In other words, Γ_{ij} contains only risks that can not be explained by Δ_{ij}, and the noise term Γ_{ij} has about the same level within each subregion $j = 1, ..., m_i$. In our application that means for example that the effect that the road density in ij has on Θ_{ij} should already be completely included in $f(\Delta_{ij})$, and that systematic differences in the error terms Γ_{ij} can appear between regions but not between subregions of the same mother region. For instance, if legal practice is not accounted for in the covariates, then it may be systematically different between Berlin and Potsdam, but there may be only unsystematic differences within Berlin and within Potsdam. If we relax assumption (3.6) to the weaker condition

$$E^{(\delta_i,\delta_{ij})}[\Gamma_{ij}] = E^{(\delta_i,\delta_{ij})}[\Gamma_i], \tag{3.8}$$

it can happen that in some subregions ij the predictor $\Theta_{ij}^{(3)}$ has a larger mean squared error than $\Theta_{ij}^{(2)}$. However, if we replace $E^{(\theta_i,\delta_i,\delta_{ij})}$ by $E^{(\delta_i,\delta_{ij})}$, we still get dominance of $\Theta_{ij}^{(3)}$ over $\Theta_{ij}^{(1)}$ and $\Theta_{ij}^{(2)}$, given that we have a sufficiently high correlation between subregions and their mother regions. Note that replacing $E^{(\theta_i,\delta_i,\delta_{ij})}$ by $E^{(\delta_i,\delta_{ij})}$ means that we average over the mean squared errors with respect to Θ_i and study them on a grainier level.

Corollary 3.3. *When replacing $E^{(\theta_i,\delta_i,\delta_{ij})}$ by $E^{(\delta_i,\delta_{ij})}$, then the first three assertions of Proposition* 3.1 *are still valid, and they are even true under the weaker assumption* (3.8).

A similar computation as in the proof of Proposition 3.1 yields the following result.

Corollary 3.4. *Under assumption* (3.8)*, it holds that*

$$\begin{aligned} &E^{(\delta_i,\delta_{ij})}\big[(\Theta_{ij} - f(\Delta_{ij}) - \lambda\Gamma_i)^2\big] \\ &= E^{(\delta_i,\delta_{ij})}\big[\Gamma_{ij}^2\big] - 2\lambda E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}] + \lambda^2\mathbb{E}^{(\delta_i,\delta_{ij})}[\Gamma_i^2]. \end{aligned} \tag{3.9}$$

The expression in the second line of (3.9) is minimized when choosing

$$\lambda = E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}]/E^{(\delta_i,\delta_{ij})}[\Gamma_i^2].$$

It may occur that the mean squared error $E^{(\delta_i,\delta_{ij})}[(\Theta_{ij} - \Theta_{ij}^{(3)})^2]$ is strictly larger than the mean squared error $E^{(\delta_i,\delta_{ij})}[(\Theta_{ij} - \Theta_{ij}^{(2)})^2]$. More precisely, this happens if and only if

$$E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}]/E^{(\delta_i,\delta_{ij})}[\Gamma_i^2] < 1/2.$$

That is typically the case when Γ_i and Γ_{ij} are weakly or negatively correlated. However, in our application to car insurance, we expect that the random variables Γ_i and Γ_{ij} are highly positively correlated, so that $\Theta_{ij}^{(3)}$ is still a better estimator than $\Theta_{ij}^{(2)}$.

In general, the quantity $E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}]/E^{(\delta_i,\delta_{ij})}[\Gamma_i^2]$ may be difficult to estimate, but we note that under the stronger assumption (3.6) we compute

$$\begin{aligned} E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}] &= E^{(\delta_i,\delta_{ij})}[\Gamma_i E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}]] \\ &= E^{(\delta_i,\delta_{ij})}[\Gamma_i^2]. \end{aligned}$$

Therefore, under (3.6), we see that the expression $E^{(\delta_i,\delta_{ij})}\big[(\Theta_{ij}-f(\Delta_{ij})-\lambda\Gamma_i)^2\big]$ is minimized for $\lambda = 1$ which is consistent to Corollary 3.2.

To put it into a nutshell, we learned that in our model the predictor $\Theta_{ij}^{(3)}$ in some sense dominates $\Theta_{ij}^{(1)}$ and $\Theta_{ij}^{(2)}$ and should be preferred.

3.3 Example

In Proposition 3.1 we provided a rigorous explanation for the improved performance of our refined predictor $\Theta_{ij}^{(3)}$ in comparison to the GDV predictor $\Theta_{ij}^{(1)}$. In this subsection we give an example of a stochastic model in which the assumptions of Proposition 3.1 are satisfied. We have already provided an economic explanation for the validity of the equation $E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}] = \theta_i - f(\delta_i)$ and now present a specific example which shows that the conditions in Proposition 3.1 can also be satisfied by a rigorously defined mathematical model.

Let $d, k \geq 1$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be an arbitrary probability space, and let $f : \mathbb{R}^d \to \mathbb{R}$ as well as $g : \mathbb{R}^d \to \mathbb{R}^k$ be measurable functions. Furthermore, for any $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$ let (W_i, W_{ij}) be a $(k+1)$-dimensional random vector with $E[W_{ij}] = 0$ and such that W_i and W_{ij} are independent. We assume that the random variables W_i, $i \in \{1, \ldots, n\}$ are identically distributed. Similarly, we also assume that the random vectors W_{ij} for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$ have the same distribution. For $\delta_i, \delta_{ij} \in \mathbb{R}^d$ we may then consider the random vector

$$(\Theta_i, \Theta_{ij}) = \left(f(\delta_i) + W_i, f\left(\delta_{ij}\right) + W_i + g(\delta_{ij})^\top W_{ij}\right), \tag{3.10}$$

so that

$$\begin{aligned}(\Gamma_i, \Gamma_{ij}) &= (\Theta_i - f(\delta_i), \Theta_{ij} - f(\delta_{ij})) \\ &= \left(W_i, W_i + g(\delta_{ij})^\top W_{ij}\right).\end{aligned}$$

In particular, $E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_i] = E^{(\theta_i,\delta_i,\delta_{ij})}[\Gamma_{ij}]$ and hence the assumption of Proposition 3.1 is satisfied.

The random variable W_i can be interpreted as regional risk effect which is not influenced by the covariates δ_i. In the subregion ij the same regional risk effect W_i should be present, but additionally it is possible to add mean-zero mixed effects $g(\delta_{ij})^\top W_{ij}$ which may depend both on covariates on the one hand and incorporate random effects on the other hand. This could be reasonable for instance if the size of a subregion would be included in the list of covariates. Indeed, we would expect the risk level in smaller regions to be more volatile than in larger regions, in the sense that its variance should be larger.

Also observe that when requiring the weaker condition (3.8) (instead of (3.6)), then it would be possible to allow mixed effects also in the region i itself. To be more precise, fix $d, k \geq 1$ and a measurable function $h : \mathbb{R}^d \to \mathbb{R}^k$. Furthermore, for each $i \in \{1, \ldots, n\}$ we consider a k-dimensional random vector W_i' such that $E[W_i'] = 0$ and such that for each $j \in \{1, \ldots, m_i\}$ the random vector (W_i, W_i') is independent of W_{ij}. We also assume that the W_i' are identically distributed for all $i \in \{1, \ldots, k\}$. Then we may consider the model where for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$ the random variables Θ_i and Θ_{ij} are defined by

$$\begin{pmatrix}\Theta_i \\ \Theta_{ij}\end{pmatrix} = \begin{pmatrix} f(\delta_i) + W_i + h(\delta_i)^\top W_i' \\ f(\delta_{ij}) + W_i + h(\delta_{ij})^\top W_i' + g(\delta_{ij})^\top W_{ij}\end{pmatrix}. \tag{3.11}$$

While the model described in (3.10) enforces a strict additive decomposition of Θ_i into the expression $f(\delta_i)$ depending only on the covariates and the random variable W_i as the regional effect not depending on the covariates, in the model given in (3.11) it is also possible to include mixed effects of the form $h(\delta_i)^\top W_i'$. We observe that as explained in Section 3.2 it is still possible to deduce

$$E^{(\delta_i,\delta_{ij})}\big[\,\Big(\Theta_{ij}-\Theta_{ij}^{(1)}\Big)^2\,\big] \geq E^{(\delta_i,\delta_{ij})}\big[\,\Big(\Theta_{ij}-\Theta_{ij}^{(3)}\Big)^2\,\big],$$

so that the refined predictor $\Theta_{ij}^{(3)}$ remains superior to $\Theta_{ij}^{(1)}$, the current standard approach in the car insurance industry. However, as we have seen in Section 3.2 the predictor $\Theta_{ij}^{(3)}$ is superior to $\Theta_{ij}^{(2)}$ if and only if $E^{(\delta_i,\delta_{ij})}[\Gamma_i\Gamma_{ij}]/E^{(\delta_i,\delta_{ij})}[\Gamma_i^2] \geq 1/2$. In our application, we expect that the risk levels Γ_i and Γ_{ij} are highly positively correlated, so that it is rather likely that the latter inequality holds.

4 Estimation of regression functions by random forests

In Section 3 we showed how information on covariates in subregions can be used to construct a refined predictor for the risk level in a subregion which exhibits a suitable optimality property. One part in the refined predictor $\Theta_{ij}^{(3)}$ introduced in (3.5) is based on the conditional expectation $f(\delta) = E[\Theta_i \mid \Delta_i = \delta]$, $\delta \in \mathbb{R}^d$, which is supposed to capture the dependence of the risk level on the covariates. In insurance practice, however, this function is also unknown and must be estimated from data. In the field of machine learning several approaches to the estimation of conditional expectations have been developed. In our application to the risk level in car insurance we are faced with a large number of possibly relevant covariates and we have to expect non-linear dependencies. A parametric linear regression approach would therefore be rather unnatural and prone to over-fitting. Therefore, we decided to follow the random-forest methodology introduced in [1] which we shall briefly recall for the convenience of the reader (see [1] and [7, Chapter 15] for further details).

A first intermediate step is the construction of so-called regression trees. A regression tree defines an approximation $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$ of the desired function f which is given by a weighted sum of indicator functions, i.e., $\widehat{f}(\delta) = \sum_{i=1}^k \alpha_i 1_{R_i}(\delta)$ for suitable values $\alpha_i \in \mathbb{R}$ and a suitable partition $(R_i)_{i\in\{1,\dots,k\}}$ of $\mathbb{R}^d$. Typically, each set R_i, $i = 1,\dots,k$ is of the form

$$R_i = \bigcap_{j=1}^{d} \left\{\delta \in \mathbb{R}^d : (-1)^{\sigma_j}\pi_j(\delta) \leq \beta_j\right\},$$

for some $\beta_j \in \mathbb{R}$, $\sigma_j \in \{0,1\}$, $j = 1,\dots,d$, where $\pi_j : \mathbb{R}^d \to \mathbb{R}$ denotes the projection to the j-th coordinate. There is no simple formula which expresses the parameters $(\alpha_i, R_i)_{i=1,\dots,k}$ in terms of given initial data $(\delta_\ell, \theta_\ell)_{\ell=1,\dots,n}$. The

$(\alpha_i, R_i)_{i=1,\dots,k}$ are thus determined by a non-trivial fitting algorithm. For the convenience of the reader, we provide a brief description of this algorithm and refer to [7, Chapter 9] for details.

The fitting proceeds in two steps, where the first step consists in defining a suitable partition $(R_i)_{i=1,\dots,k}$. For $j \in \{1,\dots,d\}$ and $s \in \mathbb{R}$ we consider the half-spaces $R_1(j,s) = \left\{\delta \in \mathbb{R}^d : \pi_j(\delta) \leq s\right\}$ and $R_2(j,s) = \left\{\delta \in \mathbb{R}^d : \pi_j(\delta) > s\right\}$. Then, we find $j \in \{1,\dots,d\}$ and $s, u_1, u_2 \in \mathbb{R}$ minimizing the expression

$$\sum_{\delta_\ell \in R_1(j,s)} (\theta_\ell - u_1)^2 + \sum_{\delta_\ell \in R_2(j,s)} (\theta_\ell - u_2)^2. \tag{4.1}$$

Stopping the fitting algorithm at this point would correspond to the estimator $\widehat{f}(\delta) = u_1 1_{R_1(j,s)}(\delta) + u_2 1_{R_2(j,s)}(\delta)$. However, typically this approximation is considered too rough and one proceeds to apply the above fitting procedure to each of the data sets $\{(\delta_\ell, \theta_\ell) : \delta_\ell \in R_1(j,s)\}$ and $\{(\delta_\ell, \theta_\ell) : \delta_\ell \in R_2(j,s)\}$. Continuing in this way, we can iteratively refine the constructed partition until a satisfactory level of approximation is attained. Typically only those members of the partition are subdivided further which contain more than a certain number of data points that has been specified in advance and the fitting algorithm terminates when each of the members in the partition contains at most this number of data points. Once the partition $(R_i)_{i=1,\dots,k}$ has been determined the coefficient α_i corresponding to R_i is defined by the average

$$\alpha_i = \left(\#\left\{(\delta_\ell, \theta_\ell) : \delta_\ell \in R_i\right\}\right)^{-1} \sum_{\delta_\ell \in R_i} \theta_\ell.$$

The algorithm described above often provides a regression tree which gives a good fit to data. However, these fits tend to be rather unstable in the sense that adding only a small number of new data points can lead to dramatic changes of the estimated function $\widehat{f}$.

Significant improvements in obtaining a stable estimate $\widehat{f}$ can be achieved by the use of *bagging*. Starting from an initial training set $M = \{(\delta_i, \theta_i)\}_{1 \leq i \leq n}$ new training sets $M_1, \dots, M_b$ can be constructed for any $b \geq 1$, where M_i is obtained from M by drawing independently n elements from M with replacement. With the help of the so-defined training sets $M_1, \dots M_b$ we get the regression trees $f_1, \dots, f_b$ which can be used to define the average $\widehat{f} = \frac{1}{b}\sum_{i=1}^{b} \widehat{f}_i$. Finally, a random forest constitutes a refinement of the bagging method, where also the subset of covariates used for fitting the regression trees is randomized. To be more precise, first fix an integer $d' \in \{1,\dots,d\}$. When constructing the regression tree $\widehat{f}_i$ from the training set M_i we use a variant of the fitting algorithm described above, where each time the refinement step (4.1) is performed the variable j is selected only from a randomly chosen subset of $\{1,\dots,d\}$ consisting of d' elements.

We conclude the present section by recalling some useful quantities related to random forest estimators. First, the so-called *out-of-bag estimator* for the mean-squared regression error may be obtained as follows. For each $(\delta, \theta) \in M$

consider the function

$$f_{\delta,\theta} = (\#\{i \in \{1,\dots,b\} : (\delta,\theta) \notin M_i\})^{-1} \sum_{\substack{i\in\{1,\dots,b\}\\(\delta,\theta)\notin M_i}} \widehat{f}_i.$$

Then the *out-of-bag estimator* ε_{oob} for the mean-squared regression error is defined as

$$\varepsilon_{oob} = \frac{1}{n} \sum_{(\delta,\theta)\in M} \left(\theta - f_{\delta,\theta}(\delta)\right)^2. \tag{4.2}$$

We note that in (4.2) it is preferable to consider the quantity $\left(\theta - f_{\delta,\theta}(\delta)\right)^2$ instead of $\left(\theta - \widehat{f}(\delta)\right)^2$, since the pair (δ,θ) was already used in the fitting of $\widehat{f}$, so that the latter alternative would lead to an estimator with significant bias.

Although the use of bagging improves prediction accuracy, it makes the random-forest methodology more difficult to interpret than models based on a single regression tree. To decrease the severeness of this disadvantage different approaches are possible. On the one hand, refined statistical approaches have been proposed in [3, 9] to reduce model complexity without deteriorating the predictive power too much. On the other hand, for random forests two kinds of variable importance scores may be computed which provide a hint as to which covariates have the largest predictive power. We only discuss the first kind of these scores in greater detail and refer to [1, 7] for further information on the second one. To determine the importance score of a variable $j \in \{1,\dots,d\}$ first fix $i \in \{1,\dots,b\}$ and consider the out-of-bag sample $M \setminus M_i$, i.e., those data points which are not used in the fitting of the i-th regression tree. We denote by $\rho : M \setminus M_i \to M \setminus M_i$, $(\delta,\theta) \mapsto (\rho(\delta),\theta)$ a function which randomly permutes the j-th coordinate. Then we compute the decrease in accuracy

$$\sum_{(\delta,\theta)\in M\setminus M_i} (f_i(\rho(\delta)) - \theta)^2 - (f_i(\delta) - \theta)^2$$

and average these quantitities over all $i \in \{1,\dots,b\}$ to obtain the importance score of type 1 for variable j.

5 Improved regional classification

Finally, we provide an application of our prediction method to real data. In our data set we consider $n = 401$ German counties. As already mentioned in Section 2, for each of these counties the GDV provides a risk level describing the relative risk in comparison to the federal average. Among all German counties, the lowest and highest observed risk levels are given by 71.15 (Elbe-Elster county) and 131.05 (Kaufbeuren county), respectively. A value of 100 corresponds to the federal average. In order to develop a suitable covariate-based predictor, we consider a vector of $d = 49$ publicly available covariates,

see e.g. [10, 11], which can be categorized into geographic data (e.g. longitude, latitude, altitude above sea level), demographic data (e.g. population density, number of registered vehicles) and road data (e.g. density of roads of a given type, density of junctions). We provide a more precise description of our data basis in the appendix.

5.1 Construction of the random-forest estimator

In the first step we construct a suitable random-forest estimate $\widehat{f}(\delta)$ of the conditional expectation $\mathbb{E}[\Theta \mid \Delta = \delta]$. Our random-forest estimate $\widehat{f}$ is based on $b = 500$ trees and for the random feature selection $d' = 15$ of the $d = 49$ covariates were used. This corresponds to the smallest out-of-bag error estimate and is in good accordance with the rule of thumb $d' \approx d/3$ suggested in [7]. Using these parameters we obtain a random-forest estimate $\widehat{f}$ for which the out-of-bag error estimate ε_{oob} can be computed as explained in (4.2). Performing this computation for our data set gives $\varepsilon_{oob} = 40.32$. Similar as in linear regression [2] we may compute an estimated R^2-value by subtracting from 1 the quotient of the out-of-bag error estimate and the empirical risk level variance, i.e.,

$$\widehat{R^2} = 1 - \frac{\varepsilon_{oob}}{\frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_i - \frac{1}{n}\sum_{j=1}^{n}\theta_j\right)}.$$

Performing this computation suggests that using the random-forest estimator a proportion of 66.63% of the risk level variance can be explained by the considered covariates. This shows that there is significant correlation between publicly available data on the one hand and the risk level of third party car insurance insurance on the other hand. However, at the same time this result also illustrates that performing a risk level prediction based only on covariates would not yield satisfactory results. In Figure 5.1 we show the importance scores of the most relevant covariates.

We see that latitude, total road density, the density of residential streets and the indicator for Bavarian counties are the four covariates with the highest importance scores (independently of the considered type of importance score). We also see that the mean altitude above sea level and the sum of county latitude and longitude seem to be of importance.

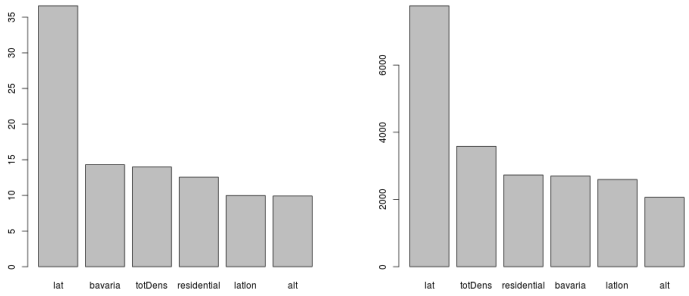

Figure 5.1: Covariates with highest importance scores (type 1/2 in the left/right hand figure), see Appendix for definition of considered variables

Furthermore, it may be interesting to perform a visual comparison of the predicted risk levels on the one hand and the actual risk levels on the other hand. In Figure 5.2 we marked those counties where the relative deviation of the predicted risk level from the actual risk level is larger than a 5% and 10% relative error threshold, respectively. Counties exhibiting an underestimated risk level are colored red, while those exhibiting an overestimated risk level are colored green. Figure 5.2 shows once more that although there is clear correlation between covariates on the one hand and risk levels on the other hand, a purely covariate-based approach might not lead to satisfactory results.

5.2 Prediction of risk level in a subregion

After having computed an estimate $\widehat{f}$ for f, see Section 5.1 we may now consider an application of the combined predictor $\Theta_{ij}^{(3)}$ introduced in (3.5) to subregions of two specific counties. As already mentioned in Section 1, when strictly adhering to the risk classification proposed by the GDV rather counter-intuitive effects can be observed at the city boundary of Berlin. While the city of Berlin is associated with the highest risk class 12, a car driver living just outside the city boundary in the adjacent Oder-Spree county only has to pay the premium associated with the lowest risk class 1. This situation is a rather absurd one and we will see that it is alleviated to a certain extent by applying our refined estimator $\Theta_{ij}^{(3)}$ to the subregion of the Oder-Spree county colored red in Figure 5.3 and to the subregion of Berlin colored green in Figure 5.3.

In Section 5.1 we computed a random-forest estimate $\widehat{f}$ for the function f based on geographic covariates, demographic covariates and covariates associated with road data. Although demographic covariates such as population density are certainly correlated with the risk level, Figure 5.1 suggests that the

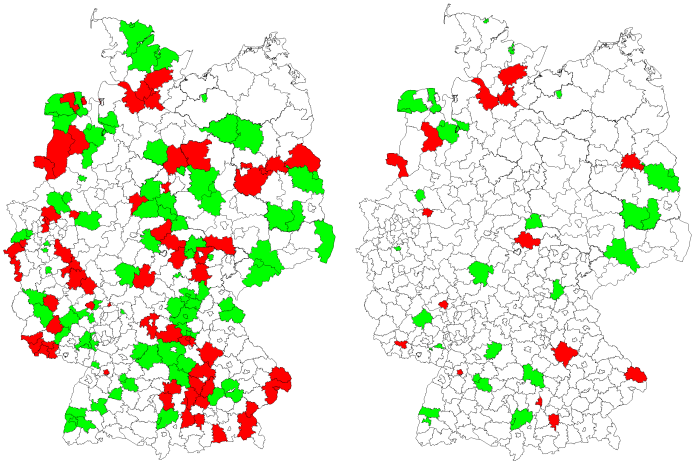

Figure 5.2: Counties with significant mispredictions (5% and 10%-threshold in the left and right hand figure, respectively)

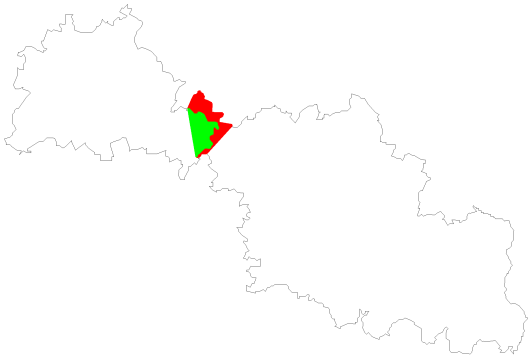

Figure 5.3: Oder-Spree county (south-east) including the considered subregion (red), as well as Berlin city (north-west) including the considered subregion (green)

corresponding road-related covariates exhibit an even better predictive power. Indeed, removing the demographic covariates from the random-forest estimator does not significantly increase the out-of-bag error estimates. Since it would require considerable effort to collect reliable demographic data for the chosen subregions, we therefore decided to omit these 13 covariates. We denote by δ_{Berlin}, $\delta_{Oder-Spree}$, $\delta_{Berlin,sub}$ and $\delta_{Oder-Spree,sub}$ the 36-dimensional vectors of covariates in Berlin city, the Oder-Spree county and the considered subregions, respectively. Evaluating the random-forest estimate $\widehat{f}$ at δ_{Berlin} and $\delta_{Berlin,sub}$, we obtain

$$\widehat{f}(\delta_{Berlin}) = 104.48 \qquad \text{and} \qquad \widehat{f}(\delta_{Berlin,sub}) = 87.54. \tag{5.1}$$

Similarly the evaluation of $\widehat{f}$ at $\delta_{Oder-Spree}$ and $\delta_{Oder-Spree,sub}$ yields

$$\widehat{f}(\delta_{Oder-Spree}) = 86.7 \qquad \text{and} \qquad \widehat{f}(\delta_{Oder-Spree,sub}) = 90.09. \tag{5.2}$$

Moreover, the risk level for the city of Berlin and the Oder-Spree county as provided by the GDV are given by

$$\theta_{Berlin} = 122.78 \qquad \text{and} \qquad \theta_{Oder-Spree} = 77.68, \tag{5.3}$$

respectively. Thus, using (5.1), (5.2) and (5.3), the corresponding values of the combined predictor $\Theta_{ij}^{(3)}$ introduced in (3.5) can be computed as

$$87.54 + (122.78 - 104.48) = 105.84.$$

for the considered subregion of Berlin

$$90.09 + (77.68 - 86.70) = 81.07.$$

for the considered subregion of the Oder-Spree county. These results are summarized in Table 1.

i	θ_i	ij	$\theta_{ij}^{(1)}$	$\theta_{ij}^{(2)}$	$\theta_{ij}^{(3)}$
Berlin	122.78	Berlin, sub	122.78	87.54	105.84
Oder-Spree	77.68	Oder-Spree, sub	77.68	90.09	81.07

Table 1: Computed values $\theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \theta_{ij}^{(3)}$ of predictors $\Theta_{ij}^{(1)}, \Theta_{ij}^{(2)}, \Theta_{ij}^{(3)}$

The values θ_{Berlin} and $\theta_{Oder-Spree}$ of the phenomenological predictors given by the GDV are far apart from each other. Likewise the values $\theta_{Berlin,sub}^{(1)}$ and $\theta_{Oder-Spree,sub}^{(1)}$ of the phenomenological predictors for the two neighboring subregions show a large difference. However, practitioners intuition suggests that the difference between the subregions should be rather small. The values $\theta_{Berlin,sub}^{(2)}$ and $\theta_{Oder-Spree,sub}^{(2)}$ of the pure explanatory predictors are very close,

but the Berlin subregion gets now a lower risk level than the Oder-Spree subregion. This change of order seems implausible and is an unwanted feature. The most convincing prediction is given by the values $\theta^{(3)}_{Berlin,sub}$ and $\theta^{(3)}_{Oder-Spree,sub}$ of the combined predictors, where the difference between the risk levels of the two neighboring subregions is significantly smaller than for the phenomenological estimator and the order of the risk levels remains unchanged.

We expect that the following refinement of our approach could yield further improvements. For practical purposes, when considering subregions close to the boundary of other subregions, it may make sense to use as a control variate not only the regression error in the given subregion, but also the regression error of the adjacent subregion. Trying to find a framework where this approach can be made more rigorous would constitute an interesting subject of further research.

Acknowledgement

C.H. has been supported by a research grant from DFG Research Training Group 1100 at Ulm University.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] L. Fahrmeir, I. Pigeot, R. Künstler, and G. Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer, Berlin, seventh edition, 2009.

[3] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2:916–954, 2008.

[4] Gesamtverband der Deutschen Versicherungswirtschaft. http://www.gdv.de/wp-content/uploads/2012/08/GDV-Grafik_Regioklasse_Haftpflicht_2013_thumb.jpg. Accessed: 22/10/2012.

[5] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.

[6] S. Gschlößl and C. Czado. Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007:202–225, 2007.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition, 2009.

[8] D.P. Kroese, T. Taimre, and Z.I. Botev. *Handbook of Monte Carlo Methods*. J. Wiley & Sons, New York, 2011.

[9] N. Meinshausen. Forest Garrote. *Electronic Journal of Statistics*, 3:1288–1304, 2009.

[10] Federal Statistical Office of Germany. https://www-genesis.destatis.de/genesis
/online/logon. Accessed: 22/10/2012.

[11] OpenStreetMap. http://www.openstreetmap.org/. Accessed: 22/10/2012.

Marcus C. Christiansen
Institute of Insurance Science
University of Ulm
89069 Ulm, Germany
e-mail: marcus.christiansen@uni-ulm.de

Christian Hirsch
Institute of Stochastics
University of Ulm
89069 Ulm, Germany
e-mail: christian.hirsch@uni-ulm.de

Volker Schmidt
Institute of Stochastics
University of Ulm
89069 Ulm, Germany
e-mail: volker.schmidt@uni-ulm.de

Appendix

For the sake of completeness we provide a description of all covariates used in the construction of our random forest estimation. These covariates can be roughly subdivided into the categories *geographic data*, *demographic data* or *road data*.

geographic data. We considered the following covariates of geographic type.

- `lat`. average latitude of the county
- `lon`. average longitude of the county
- `latlon`. sum of average longitude and average latitude
- `alt`. average altitude above sea level
- `bavaria`. indicator for being Bavarian county

demographic data. Moreover, we considered the following covariates of demographic type. We relied on publicly available data from the Federal Statistical Office of Germany [10].

- `popDens`. population density, i.e., *number of inhabitants/area*
- `pkwDens`. vehicle density, i.e., *number of automotive vehicles/area*
- `pkwRat`. ratio of numbers of vehicles and inhabitants
- `totDens`. road density, i.e., *total length of roads inside county/area*
- `totPop`. road length per inhabitant, i.e., *total length of roads inside county/number of inhabitants*
- `totPkw`. road length per vehicle, i.e., *total length of roads inside county/number of vehicles*

road data. The majority of covariates used in our application are associated with road data. Our data is taken from the *OpenStreetMap project* [11]. OpenStreetMap is a massively collaborative project to create a high-quality open-source alternative to commercially available road data. On a technical level the available data consists of an XML-database containing entries for *nodes* and *ways* connecting several nodes. The possibility to cut out subsystems of a given road system inside an arbitrarily definable boundary constitutes a feature which is especially useful for our purposes. Moreover, the data not only consists of the location of roads but also contains information regarding the type of the road (e.g. motorway, primary road, secondary road, residential street, etc.). Therefore we can consider a variety of road-related covariates such as road density, mean curvature or junction density of roads of different types. To be more precise, we include 38 further covariates of the following form. For readability we only present a selection of these covariates.

- `motorway/primar/...` road density for roads of given type, i.e., *total length of roads of given type inside county/area*
- `motorwayPop/primarPop/...` road length per inhabitant for roads of given type, i.e., *total length of roads of given type inside county/number of inhabitants*
- `aTotal`. curvature angle per *km*, i.e., *sum of absolute values of all angles occurring in polygonal representation of roads/total length of roads inside county*
- `amotorway/aprimar/...` curvature angle per *km* for roads of given type
- `juncDens`. density of junctions, i.e., *number of junctions/area*
- `mwmw/mwpr/...` density of junctions between roads of given types