Central European Economic Journal

sciendo

ISSN: 2543-6821 (online). Journal homepage: http://ceej.wne.uw.edu.pl

Google Street View image predicts car accident risk

Kinga Kita-Wojciechowska, Łukasz Kidziński

To cite this article: Kita-Wojciechowska, K., Kidziński, Ł. (2019). Google Street View image predicts car accident risk. Central European Economic Journal, 6(53), 151-163.
DOI: 10.2478/ceej-2019-0011.

To link to this article: https://doi.org/10.2478/ceej-2019-0011

Central European Economic Journal

Kinga Kita-Wojciechowska[1], Łukasz Kidziński[2]

1 Faculty of Economic Sciences, University of Warsaw, Poland, corresponding author: kingakita@gmail.com
2 Department of Bioengineering, Stanford University, Stanford, CA, USA

Google Street View image predicts car accident risk

Abstract: Road traffic injuries are a leading cause of death worldwide. Proper estimation of car accident risk is critical for the appropriate allocation of resources in healthcare, insurance, civil engineering and other industries. We show how images of houses are predictive of car accidents. We analyse 20,000 addresses of insurance company clients, collect a corresponding house image using Google Street View and annotate house features such as age, type and condition. We find that this information substantially improves car accident risk prediction compared to the state-of-the-art risk model of the insurance company and could be used for price discrimination. From this perspective, the public availability of house images raises legal and social concerns, as they can be a proxy of ethnicity, religion and other sensitive data.

Keywords: Generalized Linear Model, risk modelling, insurance pricing, satellite imagery, Google Street View

JEL Codes: G22, C83, C52

1 Introduction

Modern machine learning techniques for computer vision, like Deep Learning, provided unprecedented opportunities for academic research and industrial applications. Examples include using satellite images for deforestation monitoring in South America (Finer et al., 2018) or poverty estimation in Africa (Jean et al., 2016), prediction of skin cancer from skin lesion images (Esteva et al., 2017), or automatic detection of pulmonary tuberculosis from a chest radiograph (Lakhani & Sundaram, 2017).

One of the resources recently leveraged for research is Google Street View—a platform from Google—where images of buildings are taken using cars equipped with a set of cameras (Anguelov et al., 2010). This data source has recently been explored by researchers to answer questions in social science, for example demographic makeup of neighbourhoods across the US (Gebru et al., 2017), estimating city-level travel patterns in Great Britain (Goel et al., 2018) or crime rate in Brazil (Andersson, Birck, & Araujo, 2017).

Our work explores whether Google Street View images of houses are predictive of their residents' risk of car accident. So far, researchers were looking for determinants of car accidents among characteristics more directly related to driving, for example, driving experience (McCartt, Shabanova, & Leaf, 2003), drunk driving (Bingham, Shope, & Zhu, 2008) and using cell phones while driving (Strayer, Drews, & Crouch, 2003). There are also studies about the road and environmental conditions influencing car accidents (Karlaftis & Golias, 2002; Shankar, Mannering, & Barfield, 1995). We are not aware of any study exploring a direct link between housing conditions and car accident risk; however, a handful of research studies have proved that neighbourhood and house characteristics are correlated with health risk behaviours (Spilkova, Dzúrova, & Pitonak, 2014), which in turn correlate with driving behaviours (Rolison, Hanoch, Wood, & Liu, 2014).

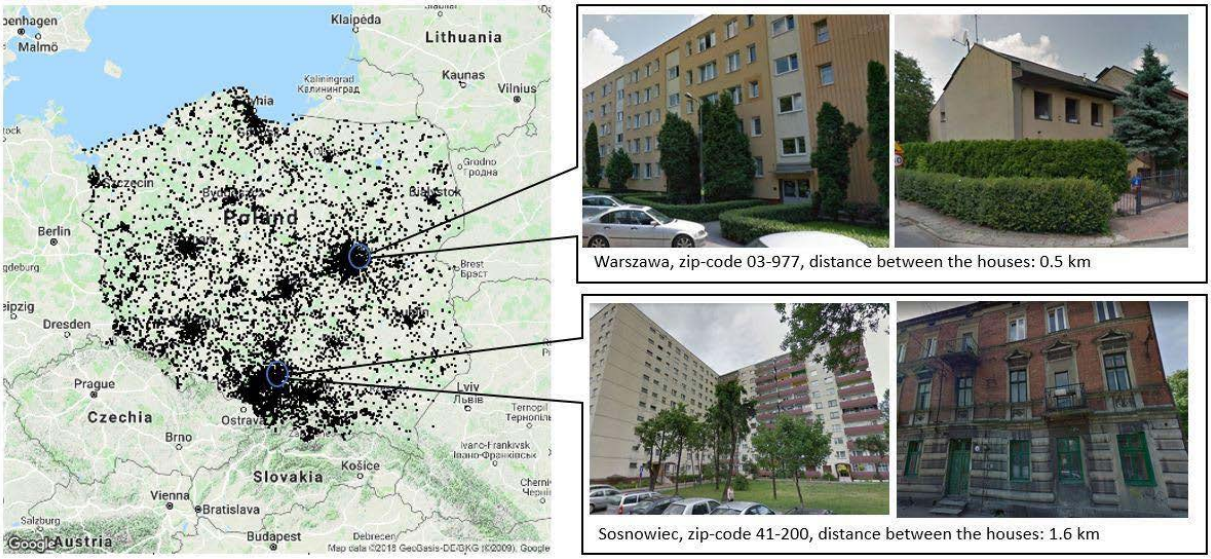

Fig. 1: Examples of extremely different houses located in the same zip code and residents of which have the same expected claim frequency by the current insurer's model.

The correlation we aim to verify in this article might be particularly interesting to the insurers. It is essential for the insurers to accurately estimate the risk of the client and set up a proper pricing in order to avoid adverse selection (Gogol, 1993). For this purpose, they search for systematic and time-invariant clients' characteristics that are observable at the moment of issuing a policy and correlate with the number of claims incurred during the insurance cover period. For example, the classical motor insurance risk factors identified worldwide are the age of the driver, the characteristics of his car, the occurrence of car accidents in the past and geography (Werner & Modlin, 2016, p.159). For this reason, the insurers tend to ask for these and other details before providing the motor insurance offer.

Although insurers often collect address information from the client, they typically use only zip-code for risk modelling and pricing purposes. Claims data aggregated to zip-codes are still too volatile and require spatial smoothing (Taylor, 2001) and further aggregation to larger geographical zones (Yao, 2008). Such a commonly used methodology is based on the assumption that neighbours are driving in a similar manner. In this article, we challenge this assumption and show that volatility can be explained at the granularity of individual addresses. Moreover, we show that this information can be extracted from publicly available images from the Google Street View (Figure 1).

Study of this insurance problem enabled following sociological and methodological discoveries: (1) features of the house correlate with the car accident risk of its resident, (2) compared to other uses of Google Street View for research, our variables are sourced from the address rather than aggregated by zip-code or district and they allow for new sociological discoveries at a very granular level, (3) variables extracted from the address (the image of a house) can be used in insurance and other industries, notably for price discrimination, (4) modern data collection and computational techniques, which allow for unprecedented exploitation of personal data, can outpace the development of legislation and raise privacy threats.

2 Results

We examine a motor insurance dataset of 20,000 records—a random sample of an insurer's portfolio collected in Poland from January 2012 to December 2015. Each record represents the characteristics of an insurance policy covering motor third party liability (MTPL) including the address of the policyholder, risk exposure defined as a fraction of the year in which the policy was active from 2013 to 2015 and the corresponding count of incurred property damage claims defined as events where any third-party property (car, motorbike, bicycle,

Fig. 2. Features annotated from Google Satellite View and Google Street View image of a particular address.

as well as fence, house, tree, etc.) has been damaged from partial or full fault of the driver of the insured car. The insurer provided us also with the expected frequency of property damage claims for those policies, estimated by their current best-in-class risk model that includes zoning based on the client's zip-code.

We collect Google Satellite View and Google Street View images for the addresses provided in the database. Six experts annotated the following features of the houses visible in the images: their type, age, condition, estimated wealth of its residents, as well as type and density of other buildings from the neighbourhood (Figure 2). Four out of six annotators gave moderately consistent answers for the common subsample of 500 addresses—Fleiss' kappa statistics indicate mostly moderate agreement among them (Table 1). These four annotators continued annotating remaining 19,371 addresses (we removed 129 addresses from the scope of this study as they were either foreign or could not be found by Google Maps), but this time each annotator was given a separate, randomly selected, set of addresses. We compared distributions of collected annotations and finally applied small corrections to match the mean and standard deviation among all four annotators.

Next, we estimated a generalized linear model (GLM) to investigate the importance of newly created variables for risk prediction (Kolyshkina, Wong, & Lim, 2004; Spedicato, Dutang, & Petrini, 2018; Werner & Modlin, 2016, p.176-183). We assume the following probabilistic model of claim frequency *f*, defined as the number of claims divided by risk exposure:

$$\log\left(\mathsf{E}\left(f\right)\right) = \log\left(\mathsf{E}\left(Y / exposure\right)\right) = \beta X$$

where Y is a number of property damage claims within MTPL insurance following Poisson distribution, X is a vector of independent variables and β is the vector of coefficients.

For relative evaluation of the value added by our approach, we introduce three models:

- Model A (null model), where vector X is $[1]$
- Model B (best-in-class insurer's model): where vector X is $[1, X_1, , X_j]$
- Model C (our model): where vector X is $\left[1, X_1, , X_j, X_{j+1}, , X_N\right]$

The insurer provided us with the realisation of the model B for each record from the dataset. That model

Tab. 1: Statistics for seven newly created variables—original granularity, inter-rater reliability of 4 selected annotators on the common set of 500 observations and significance in our risk model after applying necessary simplifications.

Variable	Original granularity	Inter-rater reliability		Risk model	
		Fleiss' kappa	Interpretation	Granularity after simplification	*p*-value
Neighbourhood type	Seven types, multi-choice	0.52	Moderate agreement	2	00.01
Building density	Scale 1–5	0.50	Moderate agreement	Not significant	
Street View quality	Good/bad/missing	0.79	Substantial agreement	2	00.02
House type	Five types, single-choice	0.69	Substantial agreement	2	00.01
House age	Scale 1–3	0.51	Moderate agreement	2	00.03
House condition	Scale 1–3	0.54	Moderate agreement	2	00.04
Wealth of residents	Scale 1–10	0.32	Fair agreement	Not significant	

was estimated on a larger undisclosed dataset and contains j predictive variables (driver characteristics, vehicle characteristics, claim history, geographical zone, etc.). Using properties of GLMs we can decompose Model C into two parts: one corresponding to Model B and one incorporating the new variables. We refer to the realisation of the Model B multiplied by exposure as an offset (Yan, Guszcza, Flynn, & Wu, 2009) and do not estimate it. Therefore, Model C takes form

$$\log\left(\mathsf{E}\left(Y\right)\right) = \beta_0 + \beta_{j+1}X_{j+1} + \ldots + \beta_N X_N + \log\left(o\,f\,f\,set\right)$$

Intuitively, in this representation, the estimated coefficients $\beta_{j+1},\ldots,\beta_N$ explain the signal that is not explained by the best-in-class risk model of the insurer (model B) and will also adjust for the earned exposure of the policy shorter than 1 year. We investigate if the values of these coefficients are non-zero, indicating that the variables we constructed provide additional predictive power to the model. We found that five out of seven newly created variables within this research were significant for predicting property damage MTPL claim frequency model (on top of many other rating variables used in the best-in-class insurer's model). We report their *p*-values in Table 1; unfortunately, the data provider did not authorise us to publish the significant levels or corresponding estimated model coefficients. Definitely more research would be needed for testing joint significance of all five selected variables, but the aim of the article is more general—to verify if there is any information in the satellite and Street View images that can be predictive of motor claim frequency and is not being captured by the existing risk factors.

To do so, we refit each of A, B and C models on an 80% train sample and check its predictive power on a 20% test sample through the corresponding Gini coefficient. We observe a significant variability of Gini coefficient on test sample—in particular for model A (null model with intercept only and no other variables selected) it varies from 20 to 38% within 20 resampling trials. We interpret it as the evidence that the dataset provided is extremely small (20,000 records) for modelling such rare events as property damage claims within MTPL insurance (average frequency of 5%).

Despite the high volatility of data, adding our five simple variables to the insurer's model improves its performance in 18 out of 20 resampling trials and the average improvement of the Gini coefficient is nearly 2 percentage points (from 38.2% to 40.1%). To put this value into perspective, the best-in-class insurer's model fitted on much bigger dataset and including a broad selection of variables (e.g. driver characteristics, car characteristics, claim history and geographical zones based on the client's zip-code) improves the Gini coefficient versus null model by 8 percentage points from ~30% to ~38% (Figure 3).

3 Discussion

We found that features visible on a picture of a house can be predictive of car accident risk, independently from classically used variables such as age, or zip code. This finding is not only a step towards more granular risk prediction models, but also illustrates a novel approach to social science, where the real-world granular data is collected and analysed at scale.

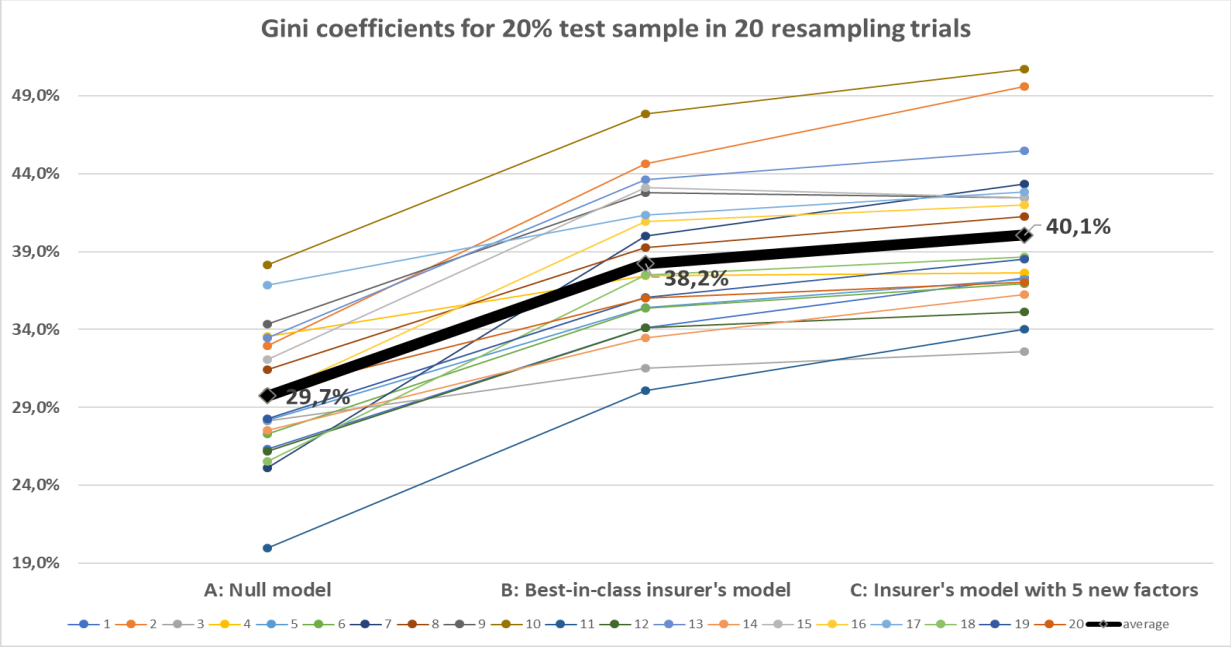

Fig. 3: Gini coefficients obtained on 20% test sample in 20 bootstrapping trials from the null model (A), the best-in-class insurer's model (B) and our model with newly created variables (C).

From the practical perspective of insurance companies, the results we present are remarkably powerful, when compared to the best-in-class insurance model. Our 5 variables, containing already some bias from the imperfect annotation, improve Gini coefficient by nearly 2 percentage points, which is massive, compared to the improvement of 8 percentage points brought by numerous variables that the insurer has already been using in his best-in-class risk model. The insurance industry could be quickly followed by the banks, as there is a proven correlation between insurance risk models and credit risk scoring (Golden, Brockett, Ai, & Kellison, 2016). The approach itself to extract valuable information from Google Street View opens a variety of opportunities not only for the financial sector. Any company that collects clients' addresses could adapt our methodology, and the deep learning technology enables to make it in an automated way on a massive scale (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014).

Such a practice, however, raises the concerns about the privacy of data stored in publicly available Google Street View, Microsoft Bing Streetside, Mapillary, or equivalent privately held datasets like CycloMedia. The consent given by the clients to the company to store their addresses does not necessarily mean a consent to store information about the appearance of their houses. In particular, features of the house may be a proxy of ethnicity, religion or other characteristics associated with a social status of a person (Braver, 2003; Gillis, 1974), which are forbidden by the law to be used for any discrimination, for example, price discrimination in certain jurisdictions (Gaulding, 1994). Fast development of modern data collection and computational techniques allows for the unprecedented exploitation of various data of clients being not even aware of it (Blitz, 2012), and the development of corresponding legislation in this matter seems to be outpaced.

The methods we present could be substantially improved by employing more annotators for the same set of the images. Potentially, the average or ensemble of their answers would match the reality better than an annotation of a single person (Levenson, Krupinski, Navarro, & Wasserman, 2015; Tran-Thanh, Stein, Rogers, & Jennings, 2014). Another limitation is the small size of the dataset provided by the insurance company, but we reduced this problem using bootstrapping and by using elementary modelling techniques, like the GLMs.

There is a question if we could extrapolate our findings on countries other than Poland. Because of historical reasons, the close neighbours in Poland may

have a very different socio-economic profile, and there is a significant heterogeneity of house types and conditions within the same zip-code. In the Western countries, the architecture might be more homogeneous; therefore, the granular information at the address level might not add much value on top of the statistics aggregated at the zip-code level.

4 Materials and methods

In this article, we describe a sample dataset obtained from the insurance company and the methodology for creating new variables from Google Street View and checking their impact on the MTPL risk prediction.

4.1 Dataset provided by the insurer

We examine a motor insurance dataset of 20,000 records—a random sample of an insurer's portfolio written in Poland from January 2012 to December 2015.

One record represents one insurance policy covering MTPL. Each record has the following characteristics attached:

- its risk exposure from 2013 to 2015 (fraction of a year that policy was active during the period 2013–2015)
- expected property damage claim frequency for that policy, estimated by the current, best-in-class risk model of the insurance company
- zip-code of the declared main driver of the car (used by the insurer to derive geographical zones)
- a set of four various addresses:

A. registered address of the policyholder
B. mail address of the policyholder
C. registered address of the car owner
D. mail address of the car owner

- the property damage claim count incurred in 2013–2015 from that policy and reported before 28 February, 2016.

Note, that there is a natural lag of reporting insurance claims to the insurer, but property damage claims from MTPL cover are rather quickly reported in Poland—95% of property damage claim are notified within first 3 months from accident occurrence, so we may assume that the observed claim count in our dataset is very close to the ultimate one.

MTPL insurance in Poland is attached to the car, not to the driver. The policy could also be purchased by a person who is neither a driver nor a car owner. Therefore, in theory, all four addresses could be different and could have a different zip-code than the one taken for the geographical zone. In practice, however, they have a lot in common:

- 84% of policies have all A, B, C and D addresses the same
- 88% of policies have common A and B addresses
- 96% of policies have common A and C addresses
- 96% of policies have common B and D addresses
- 75% of policies have common zip-code A and zip-code of the main driver
- 77% of policies have common zip-code B and zip-code of the main driver

For this study, we needed to select one address as the primary one, so we decided to select address B for the following reasons:

- The policyholder is most likely a person who is responsible for maintenance of the car and is actively using it (apart from the main driver)
- Mail address is most likely the up-to-date address of residence, while the registered address is often the one declared in the person's ID (not updated often as there is no legal obligation for it to reflect the actual residence)

On the basis of the address B, some data cleansing has been done—129 records out of 20,000 were removed from the sample as the address was either foreign or could not be found on Google Maps (Table 2).

In addition, we checked claims data for any outliers—there is only one record with three claims (where earned exposure is 0.2), and there are no records of four or more claims. Such a thin tail of our claim count distribution along with a high representation of no claim policies, and let us assume that our claims data follow Poisson distribution—a classical distribution assumed in the actuarial literature for rare events like car accidents (Goldburd, Khare, & Tevet, 2016). To confirm it, we conducted a formal test (a Chi-squared goodness-of-fit test). The test statistic X^2 is 0.08, which determines a *p*-value above 50% (from the distribution of the X^2 statistic with 1 degree of freedom). We cannot, therefore,

Tab. 2: Summary statistics of the dataset—before and after cleansing.

	Original database	After data cleansing
Number of polices	20,000	19,871
Risk exposure	11,349	11,209
MTPL PD claim count	571	570
Observed MTPL PD frequency	5,03%	5,09%

Tab. 3: Data for calculation of X^2 statistic for hypothesis verification whether claims in our dataset follow the Poisson distribution. On average $\lambda = 3.9\%$ and the corresponding $X^2 = 0.08$ with 1 degree of freedom.

Number of claims	Observed exposure (O)	Expected prob. $P(X=k)$	Expected exposure (E)	$(E-O)^2/E$
0	10,784	96%	10,785	0,00
1	417	4%	416	0,01
2	7	0%	8	0,08
All	11,209			

reject the null hypothesis that the claims follow the Poisson distribution (Table 3).

4.2 The process of creating new variables from Google Satellite View and Google Street View based on the address provided

The dataset examined in this article is a random sample of the insurer's portfolio; therefore, the geographical distribution of our addresses reflects the footprint of the insurer. It covers the whole territory of Poland with certain concentrations of policies in the big cities—Warszawa, Katowice, Kraków, Gdańsk, Szczecin, Poznań, Wrocław and Łódź (Figure 4).

For each of 19,871 addresses from the dataset, we have collected an image from Google Satellite View and an image from Google Street View (when available). We selected a random subsample of 500 addresses and asked 6 experts to annotate images from this subsample independently. They were supposed to annotate the following characteristics:

1) From Google Satellite View:
a. Types of houses and greenery prevailing in the neighbourhood (detached houses, terraced houses, blocks of flats, fields and forest)
b. Building density (on a scale 1–5)
2) From Google Street View:
a Street view quality (OK; not provided by Google; provided but its quality does not allow for annotation)
b Type of the house (detached house, terraced house, low/medium/high-rise block of flats)
c Age of the house (old, medium and new)
d The condition of the house (good, medium and bad)
e Wealth of the residents (on a scale 1–10)

Four out of six annotators gave quite consistent answers for the common subsample of 500 addresses. Fleiss' kappa statistics (Table 1) indicate mostly moderate agreement among them. We asked these four annotators to continue annotating remaining 19,371 addresses, but this time each annotator was given a separate set of addresses, not overlapping with the addresses of other annotators. After collecting all annotations, we compared the distributions of labels among annotators. Assessing the wealth of house' residents must be too subjective as its distribution varies significantly among annotators. Small differences identified in the two other variables, namely house age and house condition, were corrected by normalising the distributions among the annotators to match the mean and the standard deviation. Basic statistics of the variables after all corrections is shown in Figure 5.

It is worth noting that for 22% of addresses, there was no Google Street View available. These addresses were either in very remote locations or the road leading to them was not open to the public. Other 16% of addresses had Google Street View that did not allow for proper annotation of the house, for various reasons: the Google camera was directed at the wrong side of the road, there was an obstacle (e.g. a tree, a fence and an overtaking bus) that covered the house. As a result, only 63% of all addresses had proper Google Street View, and thus, variables, such as house type, the age of the house, condition of the house and wealth of the residents of the house, were properly annotated. Variables, such as neighbourhood type and building density, are fulfilled

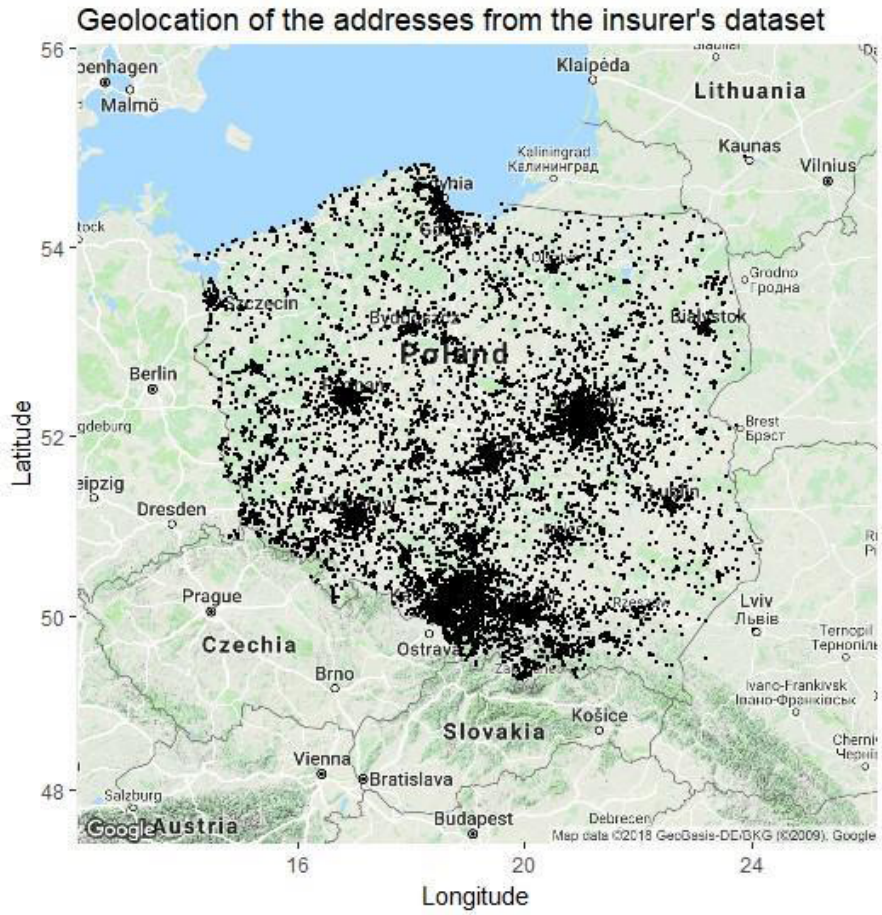

Fig. 4: Geolocation of the addresses from the dataset examined in this paper.

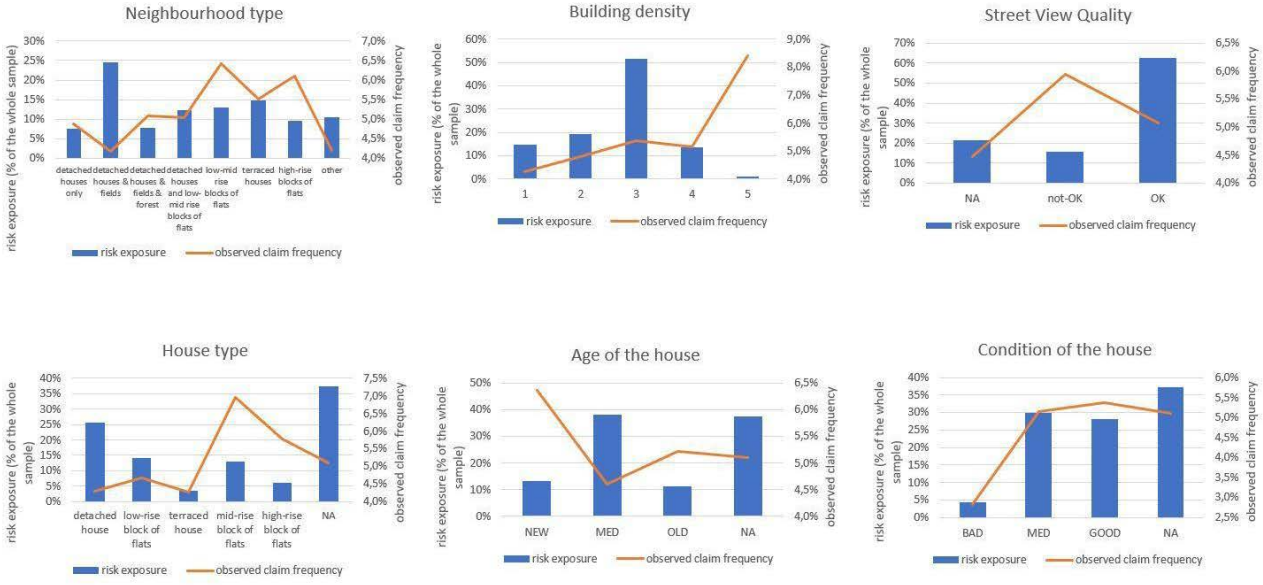

Fig. 5: Distribution of labels and corresponding observed claim frequency for the variables generated for this study.

in 100% as they are based on the Google Satellite View that was available for all observations from the dataset.

4.3 Methodology for checking the significance of the newly created variables in the risk model

In the previous section, we have presented the univariate claim frequency variable by variable (Figure 5). Some of the segments appear to have claim frequency outstanding from the population average, for example, relatively new houses, houses in bad condition, mid-rise blocks of flats, or houses surrounded by blocks of flats with no signs of greenery.

The outstanding claim frequency can be, however, driven by another variable that is already controlled by the risk model of the insurance company. For example, people living in the new houses can be relatively young, and driver's age is a classical ratemaking variable for motor insurance. There could be also some correlations among the newly created variables themselves, for example, a mid-rise block of flats is more likely to be surrounded by other blocks of flats rather than detached houses and fields. To fairly assess the impact of the newly created variables for risk prediction, we need to use a multivariate method that considers all selected variables simultaneously and automatically adjusts for exposure correlations between them.

Such a method is the GLM that has been widely adopted by the insurance pricing practitioners around the world (Cizek, Härdle, & Weron, 2005; Werner & Modlin, 2016, 176-183). GLMs extend linear models by allowing distributions of error terms other than Gaussian. In particular, residuals of models in insurance are typically assumed to follow Poisson or Gamma distributions. Despite this relaxation of assumptions on error terms, classical maximum likelihood estimates can be computed, after transforming the model with a so-called link function. Moreover, the application of log link function makes GLM coefficients interpretable and could be directly used for risk premium calculation. For these reasons, GLMs remain the most prevalent statistical tool in insurance, despite the growing popularity of complex machine learning models in other disciplines of science.

4.4 Generalized linear models

We assume following the probabilistic model of claim frequency (defined as the number of claims divided by risk exposure):

$$\log\left(\mathsf{E}\left(f\right)\right) = \log\left(\mathsf{E}\left(Y / \exp osure\right)\right) = \beta X$$

where Y is a number of property damage claims within the MTPL insurance that follows the Poisson distribution, X is a vector of independent variables and Beta's are corresponding coefficients to be estimated. Assuming N independent variables, the model formula can be written as follows:

$$\log\left(E\left(Y\right)\right) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_N X_N + \log\left(exposure\right)$$

An analogical formula is assumed for the best-in-class insurer's model, and its realisation was provided for each of the records from our dataset. We can then replace a part of the model formula by provided expected frequency that does not require estimation. Assuming the insurer uses j independent variables in the model (where $j<N$), our model formula transforms as follows:

$$\log\left(\mathsf{E}\left(Y\right)\right) = \beta_0 + \beta_{j+1} X_{j+1} + \ldots + \beta_N X_N + \log\left(o\,f\,f\,set\right)$$

We estimate such a model formula in R package. The variables are being added to the model step by step, and the necessary grouping of levels is being made meanwhile to achieve the most robust results. The modelling process is iterated until all factors used in the model appear significant (p-value < 0.05).

4.5 Model evaluation

Once the modelling process is finished, we validate the model by refitting it on 80% train sample and checking its performance on 20% test sample through the Gini coefficient. Gini coefficient is most commonly known as a measure of the inequality of income, but it has been adopted by insurance practitioners as a metric for model validation and model comparison (Frees, Meyers, & Cummings, 2011). It is computed as follows:

1. The policies in 20% of test sample are sorted from the lowest to the highest claim frequency expected by the model fitted on the 80% of train sample

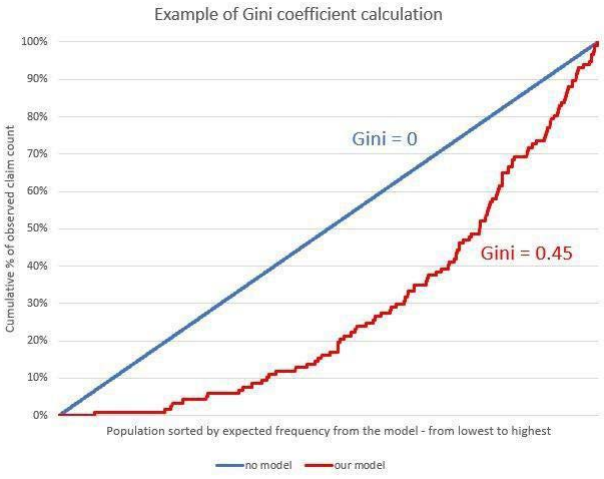

Fig. 6: Illustration of the Gini coefficient computation for one of the bootstrapping trials.

2. The cumulative observed claim count from sorted policies in 20% of test sample is plotted on the graph (representing inequality of risk distribution in the population, analogically to inequality of wealth distribution in Lorenz curve (Lorenz, 1905))
3. Gini coefficient is computed as the area between the Lorenz curve and the no-discrimination line multiplied by 2 (where the Lorenz curve is described in point 2 and illustrated in Figure 6) (Gini, 1921)

Our preliminary analysis has shown the variability of the Gini coefficient due to the small size of the dataset provided. To reduce this variability in the analysis of model performance, we compute the estimates of the Gini coefficient from 20 resampling trials, each time randomly assigning observations to train and test set from the beginning.

Acknowlegments

We thank Michał Skwarek for assistance in data preparation and his valuable insights that greatly improved the research.

References

[1] Andersson, V. O., Birck, M. A. F., & Araujo, R. M. (2017). Investigating crime rate prediction using street-level images and Siamese convolutional neural networks. In E. Teles & C. Brackmann (Eds.), *Computational neuroscience* (pp. 81–93). Cham, Switzerland: Springer International Publishing.

[2] Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., … Weaver, J. (2010). Google street view: Capturing the world at street level. *Computer*, *43*(6), 32–38.

[3] Bingham, C. R., Shope, J. T., & Zhu, J. (2008). Substance-involved driving: Predicting driving after using alcohol, marijuana, and other drugs. *Traffic Injury Prevention*, *9*(6), 515–526.

[4] Blitz, M. J. (2012). The right to map (and avoid being mapped): Reconceiving first amendment protection for information-gathering in the age of Google Earth. *The Columbia Science and Technology Law Review*, *14*, 115.

[5] Braver, E. R. (2003). Race, Hispanic origin, and socioeconomic status in relation to motor vehicle occupant death rates and risk factors among adults. *Accident; Analysis and Prevention*, *35*(3), 295–309.

[6] Cizek, P., Härdle, W. K., & Weron, R. (2005). *Statistical tools for finance and insurance*. Berlin, German: Springer Science & Business Media.

[7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

[8] Finer, M., Novoa, S., Weisse, M. J., Petersen, R., Mascaro, J., Souto, T., … Martinez, R. G. (2018). Combating deforestation: From satellite to intervention. *Science*, *360*(6395), 1303–1305.

[9] Frees, E. W., Meyers, G., & Cummings, A. D. (2011). Summarizing insurance scores using a Gini Index. *Journal of the American Statistical Association*, *106*(495), 1085–1098.

[10] Gaulding, J. (1994). Race sex and genetic discrimination in insurance: What's fair. *Cornell Law Review*, *80*, 1646.

[11] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the

demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(50), 13108–13113.

[12] Gogol, F. (1993). The Value of Information in Insurance Pricing. *The Journal of Risk and Insurance*, 60(1), 119-128.

[13] Gillis, A. R. (1974). Population density and social pathology: The case of building type, social allowance and juvenile delinquency. *Social Forces; a Scientific Medium of Social Study and Interpretation*, *53*(2), 306–314.

[14] Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal of Nepal*, *31*(121), 124–126.

[15] Goel, R., Garcia, L. M. T., Goodman, A., Johnson, R., Aldred, R., Murugesan, M., … Woodcock, J. (2018). Estimating city-level travel patterns using street imagery: A case study of using Google Street View in Britain. *PloS One*, *13*(5), e0196521.

[16] Goldburd, M., Khare, A., & Tevet, C. D. (2016). Generalized linear models for insurance rating. In *Casualty Actuarial Society*. Retrieved from https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf.

[17] Golden, L. L., Brockett, P. L., Ai, J., & Kellison, B. (2016). Empirical evidence on the use of credit scoring for predicting insurance losses with psycho-social and biochemical explanations. *North American Actuarial Journal: NAAJ*, *20*(3), 233–251.

[18] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

[19] Karlaftis, M. G., & Golias, I. (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident; Analysis and Prevention*, *34*(3), 357–365.

[20] Kolyshkina, I., Wong, S., & Lim, S. (2004). Enhancing generalised linear models with data mining. In *Casualty Actuarial Society* (pp. 279–290).

[21] Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284*(2), 574–582.

[22] Levenson, R. M., Krupinski, E. A., Navarro, V. M., & Wasserman, E. A. (2015). Pigeons (*Columba livia*) as trainable observers of pathology and radiology breast cancer images. *PloS One*, *10*(11), e0141357.

[23] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

[24] McCartt, A. T., Shabanova, V. I., & Leaf, W. A. (2003). Driving experience, crashes and traffic citations of teenage beginning drivers. *Accident; Analysis and Prevention*, *35*(3), 311–320.

[25] Rolison, J. J., Hanoch, Y., Wood, S., & Liu, P.-J. (2014). Risk-taking differences across the adult life span: A question of age and domain. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *69*(6), 870–880.

[26] Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident; Analysis and Prevention*, *27*(3), 371–389.

[27] Spedicato, G. A., Dutang, C., & Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. *Variance: Advancing the Science of Risk*, *111*(2), 69-89.

[28] Spilkova, J., Dzúrova, D., & Pitonak, M. (2014). Perception of neighborhood environment and health risk behaviors in Prague's teenagers: A pilot study in a post-communist city. *International Journal of Health Geographics*, *13*, 41.

[29] Strayer, D. L., Drews, F. A., & Crouch, D. J. (2003). Fatal distraction? A comparison of the cell-phone driver and the drunk driver. In *Driving Assessment Conference* (Vol. 2). University of Iowa. doi:10.17077/drivingassessment.1085.

[30] Taylor, G. (2001). Geographic premium rating by whittaker spatial smoothing. *ASTIN Bulletin: The Journal of the IAA*, *31*(1), 147–160.

[31] Tran-Thanh, L., Stein, S., Rogers, A., & Jennings, N. R. (2014). Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, *214*, 89–111.

[32] Werner, G., & Modlin, C. (2016). *Basic ratemaking* (5 ed.). Casualty Actuarial Society.

[33] Yan, J., Guszcza, J., Flynn, M., & Wu, C.-S. P. (2009). Applications of the offset in property-casualty

predictive modeling. In *Casualty Actuarial Society E-Forum, Winter 2009* (p. 366).

[34] Yao, J. (2008). Clustering in ratemaking: Applications in territories clustering. *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society-Arlington, Virginia*, 170–192.

[35] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 487–495). Red Hook, NY: Curran Associates.