Simulation Engine for Adaptive Telematics data

Abstract

This article provides a simulation engine, which we call `SEAT` (Simulation Engine for Adaptive Telematics), for the flexible generation of an insurance claims dataset with driver telematics information that matches the specific profile of a target market.

Generation of an adaptive telematics data via `SEAT` follows the following two-stage process. In the first stage, `SEAT` uses pre-determined distributions of some traditional policy characteristics of the target market as inputs and replicates these policy characteristics based on their distributions. Afterwards, `SEAT` generates the rest traditional, telematics variables and insurance claims accordingly given configurations of the (traditional) policy characteristics with possible perturbations.

As a specific example, we illustrate an adaptive telematics dataset to match South Korean insurance market and compare its behavior to the source telematics dataset where the algorithm is based on. We hope that both the practitioners and researchers utilize this publicly available simulation engine (https://github.com/bheeso/SEAT.git) and an adaptive dataset to explore usefulness of the driver telematics data to develop diverse models for the usage-based insurance.

Keywords: Data mining, synthetic data, adaptive data generation, telematics

1 Introduction and motivation

Due to recent innovations in technology, it became feasible to keep track of an automobile use that includes rich information such as mileage, speed, acceleration with telematics devices. While 'telematics' refers to "the use or study of technology that allows for information to be sent over long distances using computers" (Oxford dictonary) so that it has been considered in a variety of fields. More specifially, insurance industry has been applying telematics in practice, for example, usage-based insurance (UBI) to exploit available information from the record of driving on top of traditionally observable policy and driver characteristics. By installing a plug-in device or using a mobile application, the insurers can easily collect relevant driving information of a specific policyholder, which can be potentially used to classify risk in a more sophisticated way.

Inspired by such interest in practice, use of telematics for automobile insurance has been actively studied in the actuarial literature. Ayuso et al. (2014) is one of the earliest paper that addressed

possible use of telematics for Pay-as-you-Drive insurance. They showed that vehicle usage between novice and experienced young drivers are quite different by analyzing the telematics data.

There are some research articles that showed telematics data can provide richer information on top of the traditional rating variables. According to Ayuso et al. (2016), gender, a traditionally and widely used rating variable, is not a significant rating factor once it is controlled by the observed driving record. It has been found that there is significant effect of driving habits on the expected number of claim by Ayuso et al. (2019). Guillen et al. (2019) showed that use of telematics devices is helpful to predict the excess of zero in claim frequency more precisely. One can find more examples on the use of telematics information to improve traditional risk classification models, which includes but not limited to Boucher et al. (2017), Gao et al. (2019), Pesantez-Narvaez et al. (2019), and Pérez-Marín et al. (2019). Recently, use of telematics data also has been studied to discover near-miss events. (Guillen et al., 2020, 2021)

In spite of the presence of active research both in academica and the industry, however, access to telematics data has been quite limited due to possible privacy issues so that it has been hard to try more diverse methods for telematics data analytics for most of members in the actuarial and insurance community. In this regard, So et al. (2021) proposed a novel algorithm to replicate an actual insurance portfolio with telematics data and synthesize a publicly available dataset. However, the synthetic dataset based on the algorithm of So et al. (2021) replicates the distribution of independent and dependent variables almost identically, which might make use of the synthetic dataset less practical if the characteristics of the interested portfolio is quite different from the original telematics dataset. Therefore, we propose an advanced algorithm of adaptive telematics data generation so that one can incorporate prior knowledge on the specific characteristics of an insurance market or a company.

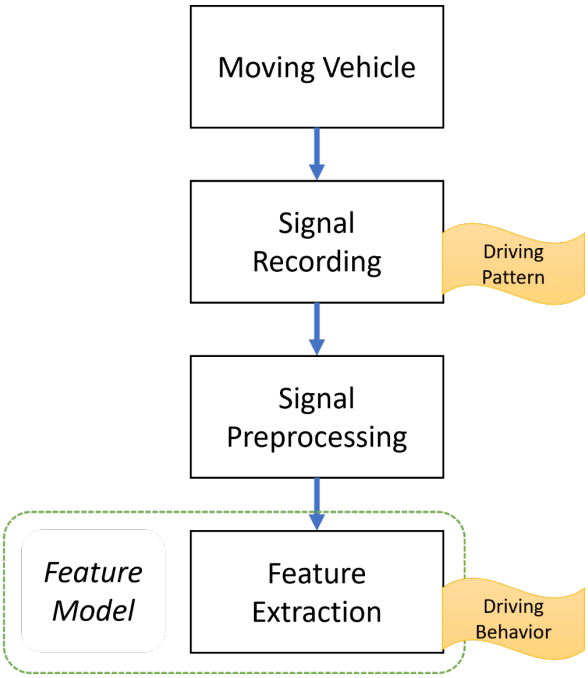
The remaining part of this paper is organized as follows. In Section 2, we introduce general concept and characteristics of telematics information and synthetic datasets. In Section 3, we provide our main result, construction of Simulation Engine for Adaptive Telematics data (`SEAT`) in detail with relevant theoretical foundations and actual implementation. In Section 4, we illustrate an empirical application of the proposed method by generating a portfolio that matches South Korean insurance market. We also discuss the characteristics of the generated portfolio in comparion to the original portfolio. We conclude this paper with some remarks in Section 5.

2 Telematics and synthetic data

As mentioned in Section 1, telematics data on vehicle use are collected and transmitted on a real time basis so that the aspects of data collection and processing of telematics data are quite different from those of the so-called traditional data. According to Johanson et al. (2014), the total size of the driving signal dataset per each day will be about 560 gigabytes for 1,000 vehicles. Considering usual automobile insurance portfolio sizes that include millions of drivers, the following issues;

effectiveness in recording, data privacy (Duri et al., 2002, 2004), and feature extraction need to be considered carefully.

Figure 1: Feature extraction from telematics data (Weidner et al., 2016)



Conventionally, telematics data has been feature engineered into specific forms of attributes that can be meaningful and ready to use for prediction of auto insurance claims. Table 1 introduces a brief list of summary statistics that can be used in practice (Gerardo and Lee, 2009).

Table 1: Examples of extracted features from telematics data

Category	GPS	Mileage	Operation
Items	Car location	Weekend mileage	Average speed
	Running status	Night mileage	Average acceleration
	Travel time	Average monthly mileage	Average RPM

Recently, Wüthrich (2017) and Gao and Wüthrich (2018) suggested a different framework of feature extraction and dimension reduction of telematics data. Since telematics data inherently form a continuum of observations, they considered the velocity-acceleration heatmaps and analyzed them via a K-means clustering algorithm for risk classification of car drivers. Henceforth, we will focus on the emulation and analysis for a summarized form of telematics data rather than the raw data itself, as in most of the research articles in the actuarial literature.

Even in a summarized form, however, it has been extremely difficult to access a publicly available telematics dataset for the actuarial and insurance commnuity, mainly due to concerns of privacy that make many insurers hesitate to provide their data to researchers. In this regard, there has been keen interest on synthetic data, which is free of privacy issues but still maintains essential characteristics

of the original dataset and easily used for construction and validation of statistical models for improved risk classifications. For example, Gan and Valdez (2018) provided a synthetic dataset of a large variable annuity portfolio that can be used for development of annuity valuation or hedging techiniques such as metamodeling. Gabrielli and Wüthrich (2018) proposed an individual claims history simulation machine that helps researchers to calibrate their own individual or aggregate reserving models. Cote et al. (2020) applied generative adversarial networks (GAN) to synthesize a property & casualty ratemaking dataset, which can be potentially used for predictive analytics in the ratemaking purpose. Avanzi et al. (2021) introduced an individual insurance claim simulator with feature control, which enables the modelers to explore validity of their reserving methods by back-testing.

To the best of authors' knowledge, So et al. (2021) is the first research article that addressed synthesization of a dataset that includes features engineered from an actual telematics dataset. They used a three-stage process that utilizes various machine learning algorithms. Firstly, they generated feature variables of a synthetic portfolio via an extended version of Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) algorithm. Secondly, the corresponding number of claims were simulated via binary classifications using feedforward neural networks (FFN). Lastly, the corresponding aggregated amount of claims were simulated via regression FFN.

Despite the novelty and originality, use of extended SMOTE algorithm for feature generation would end up with almost identical distributions of the traditional and telematics feature variables of the synthetic and original datasets, as shown in Appendix of So et al. (2021). It could be a possible concern for researchers and especially practitioners that want to use a synthetic dataset due to heterogeneity of a feature portfolio in between the target market and the original data. Therefore, we propose an advanced version of algorithm in the following section that uses prior information of a target market as inputs and enables to synthesize a feature portfolio similar to the one from the target market.

3 Simulation of adaptive telematics data

3.1 Desirable characteristics of a synthetic telematics dataset

Here we discuss some desirable characteristics of a synthetic telematics dataset before the proposed method, `SEAT`, is introduced in detail. Firstly, it should be open to public. As mentioned in the previous sections, limited access to the telematics data has been one of the most serious obstacles to develop and back-test diverse methods of ratemaking with telematics features. Secondly, it should be flexible enough to satisfy the needs of the modelers with specific interests. Lastly, granularity of the dataset should be assured in order that a predictive model with various features can be trained, tested, and applied in individual risk classifications with given synthetic dataset.

In this regard, we targeted to develop a method that is available and open to public so that one can access the sample dataset, data generation routine, and the source codes. Further, the resulting

dataset is fully granular and also flexible in the sense that it can match various target market profiles of interest. While one can consider more desirable characteristics, such as longitudinality and multivariate claims to capture possible serial and/or between coverage dependence, we defer these topics to future works.

3.2 Description of the source data

To introduce the construction and use of `SEAT`, here we use a synthetic insurance claim data as a source data, which consists of traditional and telematics features, and two response variables. This synthetic data is introduced in So et al. (2021) and publicly available from http://www2.math.uconn.edu/~valdez/data.html (accessed on 15 June 2022).

There are 52 variables in total in the source dataset, which belong to the following three main categories:

- 11 traditional features, such as insurance exposure, age of the insured, and main use of vehicle,
- 39 telematics features, including but not limited to total distance driven, number of sudden acceleration and brakes, and
- two response variables describing the claim frequency and aggregated claim amounts.

Table 2 provides the name, description, and data attributes of the available features in the dataset. Note that attributes of the features are important since they affect the data generation scheme with random perturbation, as introduced in Step 4 of Section 3.2. For detailed information and preliminary analysis of the dataset, see Section 3 of So et al. (2021).

3.3 Adaptive data generation scheme

In this section, we provide details of the `SEAT` algorithm. As mentioned earlier, the proposed algorithm uses pre-determined distributions of the traditional covariates such as gender, residence, age, and main purpose of use of the insured vehicle, which are the information easily accessible in the market. In our source dataset, they are named as `Insured.sex`, `Region`, `Insured.age`, and `Car.use`, respectively. Putting this prior knowledge into the `SEAT` algorithm as inputs, one can adapt an original feature portfolio to the one for the target market.

Furthermore, the proposed `SEAT` algorithm is readily available with the following link (https://github.com/bheeso/SEAT.git) and enables the insurance practitioners and researchers to get an adaptive data to fit their need for a feature portfolio.

Here we assume that we know the benchmark ratios of classes in four covariates (`Insure.age`, `Insure.sex`, `Region`, `Car.use`) and set these ratios as the inputs of the algorithm. Table 3 shows how the inputs are defined. For examples, $P^*(A1)$ (in Table 3) indicates the ratio of the insureds aged between 16 and 30. There are in total 13 inputs. By applying the `SEAT` algorithm, original portfolio in source data will adapt to a portfolio having these ratios for four variables.

Type	Variable	Description	Attributes
Traditional	`Duration`	Duration of the insurance coverage of a given policy, in days	Ordinal
	`Insured.age`	Age of insured driver, in years	Ordinal
	`Insured.sex`	Sex of insured driver (Male/Female)	Categorical
	`Car.age`	Age of vehicle, in years	Ordinal
	`Marital`	Marital status (Single/Married)	Categorical
	`Car.use`	Use of vehicle: Private, Commute, Farmer, Commercial	Categorical
	`Credit.score`	Credit score of insured driver	Ordinal
	`Region`	Type of region where driver lives: rural, urban	Categorical
	`Annual.miles.drive`	Annual miles expected to be driven declared by driver	Continuous
	`Years.noclaims`	Number of years without any claims	Ordinal
	`Territory`	Territorial location of vehicle	Categorical
Telematics	`Annual.pct.driven`	Annualized percentage of time on the road	Continuous
	`Total.miles.driven`	Total distance driven in miles	Continuous
	`Pct.drive.xxx`	Percent of driving day xxx of the week: mon/tue/.../sun	Continuous
	`Pct.drive.xhrs`	Percent vehicle driven within x hrs: 2hrs/3hrs/4hrs	Continuous
	`Pct.drive.xxx`	Percent vehicle driven during xxx: wkday/wkend	Continuous
	`Pct.drive.rushxx`	Percent of driving during xx rush hours: am/pm	Continuous
	`Avgdays.week`	Mean number of days used per week	Continuous
	`Accel.xxmiles`	Number of sudden acceleration 6/8/9/.../14 mph/s per 1000miles	Ordinal
	`Brake.xxmiles`	Number of sudden brakes 6/8/9/.../14 mph/s per 1000miles	Ordinal
	`Left.turn.intensityxx`	Number of left turn per 1000miles with intensity 08/09/10/11/12	Ordinal
	`Right.turn.intensityxx`	Number of right turn per 1000miles with intensity 08/09/10/11/12	Ordinal
Response	`NB_Claim`	Number of claims during observation	Ordinal
	`AMT_Claim`	Aggregated amount of claims during observation	Continuous

Table 2: Variable names and descriptions of the source dataset

Variable	Classes	Inputs
Insured.age	[16,30) / [30,40) / [40,50) / [50,60) / [60,103]	$P^*(A1)$, $P^*(A2)$, $P^*(A3)$, $P^*(A4)$, $P^*(A5)$
Insured.sex	Male / Female	$P^*(M)$, $P^*(F)$
Region	Rural / Urban	$P^*(R)$, $P^*(U)$
Car.use	Private / Commute / Farmer / Commercial	$P^*(C1)$, $P^*(C2)$, $P^*(C3)$, $P^*(C4)$

Table 3: Description of the inputs associated with the four variables

When the input values are given, the following algorithm is suggested in order to generate an adaptive data from the source data:

- Step 1: Calculate conditional distributions of four covariates (See Figure 2) using source data and inputs.

  In the algorithm, since there may exist collinearity among the four variables, we use modified conditional distributions to reflect such collinearities. To obtain the modified conditional distributions, we use the following probability rules. (For simplicity, we write $P^*(\texttt{Variable 1} = x)$, $P^*(\texttt{Variable 2} = y \,|\, \texttt{Variable 1} = x)$, and $P^*(\{\texttt{Variable 1} = x\} \cap \{\texttt{Variable 2} = y\})$ as $P^*(x)$, $P^*(y|x)$, and $P^*(x \cap y)$, respectively unless there is a room for confusion.)

$$P^*(R|F) = \frac{P^*(R \cap F)}{P^*(R \cap F) + P^*(U \cap F)} \text{ and } P^*(R|M) = \frac{P^*(R \cap M)}{P^*(R \cap M) + P^*(U \cap M)},$$

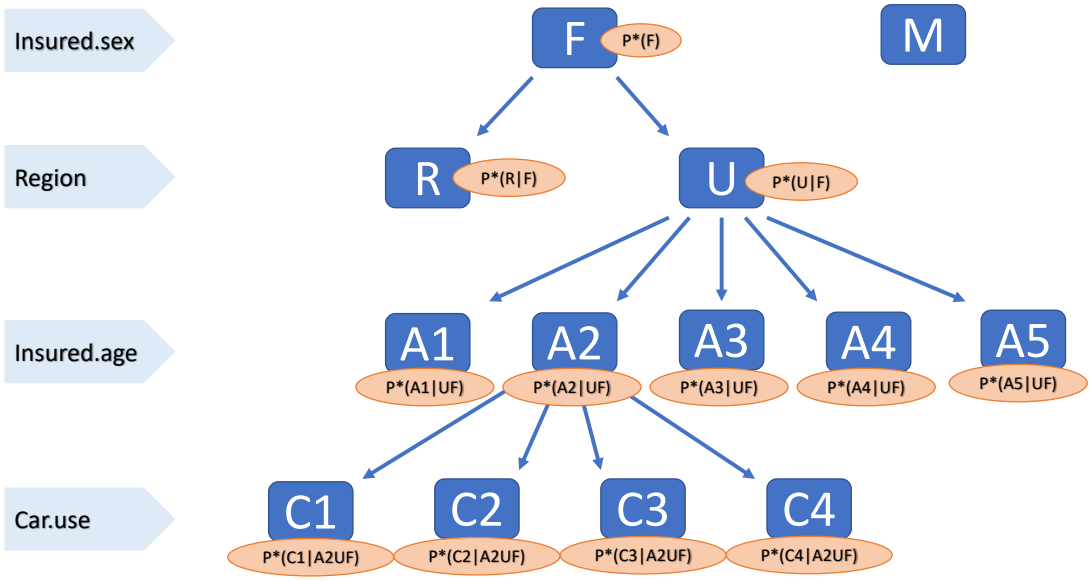


Figure 2: A branch of generating conditional distributions with four covariates

where

$$\begin{aligned} P^*(R\cap F) &= P(R\cap F)\frac{P^*(R)}{P(R)},\ P^*(U\cap F) = P(U\cap F)\frac{P^*(U)}{P(U)},\\ P^*(R\cap M) &= P(R\cap M)\frac{P^*(R)}{P(R)},\ P^*(U\cap M) = P(U\cap M)\frac{P^*(U)}{P(U)}, \end{aligned} \tag{1}$$

and $P(\cdot)$ means the sample ratio calculated from the source data. One can easily check that (1) is equivalent to

$$P^*(F|R) = P(F|R) \text{ and } P^*(F|U) = P(F|U).$$

We may call $\frac{P^*(R)}{P(R)}$ and $\frac{P^*(U)}{P(U)}$ as adjustment factors since they play a role in adjusting ratios inside of the original portfolio to the ones in adaptive portfolio. We get these adjustment factors from the following equations:

$$P^*(R\cap F) + P^*(R\cap M) = P^*(R),$$

$$P^*(R\cap F) = P(F|R)P^*(R) = P(R\cap F)\frac{P^*(R)}{P(R)},$$

$$P^*(R\cap M) = P(M|R)P^*(R) = P(R\cap M)\frac{P^*(R)}{P(R)}.$$

Other conditional ratios are subsequently calculated in the same way.

- Step 2: Create a random sample from the standard uniform distribution and categorize it based on each conditional distribution in the order of `Insured.sex`, `Region`, `Insured.age`, and `Car.use`. Firstly, one can sample a random observation `Insured.sex` from its benchmark distribution [P^*(`Insured.sex` $= F$) and P^*(`Insured.sex` $= M$)]. If a uniform random number is categorized into female (`F`), then, in next stage, a newly generated uniform random number is categorized based on conditional distribution of `Insured.sex` [P^*(`Region` $= R \,|\, $ `Insured.sex` $= F$) and P^*(`Region` $= U \,|\, $ `Insured.sex` $= F$)].

- Step 3: By repeating Step 2 N times, gain N configurations of the four (traditional) covariates $\{\mathcal{C}_i\}_{i=1,\ldots,N}$ where N is the number of total observations in the adaptive portfolio to be generated.

- Step 4: From the original portfolio, sample N observations of the remaining variables (telematics and claims information) from their empirical distributions given each configuration with white Gaussian noises. For example, if a configuration is $\mathcal{C}_i = \{$`Car.use` $= F$, `Region` $= R$, `Insured.age` $= A1$, `Insured.sex` $= C1\}$, then the corresponding telematics information $\mathcal{T}_i = (T_i^{(1)}, \ldots, T_i^{(p_\mathcal{T})})$, number of claims N_i, and total amounts of claims S_i are generated as follows:

$$\begin{aligned} T_i^{(j)} &\simeq \tilde{F}^{-1}_{T^{(j)}}(U_i|\mathcal{C}_i) + \sigma Z_i, \\ N_i &\simeq \tilde{F}^{-1}_{\mathcal{N}}(U_i|\mathcal{C}_i) + \sigma Z_i, \\ S_i &= \tilde{F}^{-1}_{\mathcal{S}}(U_i|\mathcal{C}_i) + \sigma Z_i, \end{aligned}$$

  where Z_i is a random sample from the standard normal distribution, $p_\mathcal{T}$ is the number of telematics features, $\tilde{F}_X(\cdot|\mathcal{C})$ is the empirical distribution of feature X given the configuration of traditional covariates $\mathcal{C}$, and σ is the input that controls the degree of random perturbation. Note that for the ordinal variables (for example, number of claims or number of sudden brakes), $T_i^{(j)}$ and/or N_i are rounded to the nearest ordinal numbers, respectively.

Algorithm 1 summarizes the steps for data generation given input values. Note that the proposed algorithm provides a room for flexibility by (sub)data generation. For example, if a modeler is interested in analyzing the policyholders in urban areas, then one can impose $P^*(R) = 0$ so that the resulting dataset can only contain the policyholders in urban areas and their corresponding features.

Algorithm 1: `Simulation Engine for Adaptive Telematics (SEAT)`

Input: $[P^*(M),\ P^*(F)]$,
$[P^*(R),\ P^*(U)]$,
$[P^*(A1),\ P^*(A2),\ P^*(A3),\ P^*(A4),\ P^*(A5)]$,
$[P^*(C1),\ P^*(C2),\ P^*(C3),\ P^*(C4)]$
$N,\ \sigma$

Output: adaptive dataset to target feature portfolio (size is N)

Import the data introduced in So et al. (2021) as source dataset;

Calculated conditional distributions, $P^*(R|F)$, $P^*(U|F)$, $P^*(R|M)$, $P^*(U|M)$, $P^*(A1|RF)$, $P^*(A2|RF)$, $\cdots$, $P^*(A5|UM)$, $P^*(C1|A1RF)$, $\cdots$, $P^*(C4|A5UM)$;

for $i = 1, \ldots, N$ **do**

Create a random sample, u_i, from $U(0,1)$;

Categorize u_i based on conditional distributions for `Insured.sex`, `Region`, `Insured.age`, and `Car.use`, subsequently, and create a configuration, $\mathcal{C}_i$;

From source data, randomly sample an observations having $\mathcal{C}_i$ with white Gaussian noises (σ) as follows:

$$T_i^{(j)} \simeq \tilde{F}^{-1}_{T^{(j)}}(U_i|\mathcal{C}_i) + \sigma Z_i,$$
$$N_i \simeq \tilde{F}^{-1}_{\mathcal{N}}(U_i|\mathcal{C}_i) + \sigma Z_i,$$
$$S_i = \tilde{F}^{-1}_{\mathcal{S}}(U_i|\mathcal{C}_i) + \sigma Z_i$$

end

Return the adaptive dataset: $\{(\mathcal{C}_1,T_1,N_1,S_1),\ (\mathcal{C}_2,T_2,N_2,S_2),\ \cdots,\ (\mathcal{C}_N,T_N,N_N,S_N)\}$;

4 Empirical application: Synthetized telematics data for South Korean insurance market

In this section, we provide a specific example of synthetized telematics dataset that is tailored to South Korean insurance market with pre-determined inputs, and analyze their properties in comparison to the original telematics data. While we use South Korean insurance market as a specific example, we also want to emphasize that the proposed algorithm could be applied to any target market, in national, regional, industry, or company level depending on the specific interest of a modeler.

South Korean insurance market has grown up rapidly and ranked the 7th in terms of total premium volume in 2020 (Korean Insurance Research Institute, 2021), which consists of 3.1% of the global market share. Nevertheless, uses of telematics data both in actuarial practice and research are still in a developing stage in South Korea. While Han (2016) discussed regulatory and legal issues for the use of telematics data at usage-based insurance in South Korean insurance market, there was lack of follow-up research on the implementation of ratemaking methods with telematics data in South Korean insurance market, partially due to lack of publicly available telematics data. Further,

there is only one insurance company, Carrot Insurance, which actively utilizes driver telematics information in their ratemaking scheme. In this regard, we extract basic profiles of South Korean insurance market from various sources (The Korean National Police Agency, 2020; Korean Statistical Information Service, 2020a; The Korean Ministry of Land, Infrastructure and Transport, 2020; Korean Statistical Information Service, 2020b) and use these as the inputs for the adaptive portfolio generation. Table 4 provides the specification of inputs for a tailored data generation.

Variable	Inputs
Insured.age	$P^*(A1) = 0.16$, $P^*(A2) = 0.21$, $P^*(A3) = 0.24$, $P^*(A4) = 0.22$, $P^*(A5) = 0.17$
Insured.sex	$P^*(M) = 0.6$, $P^*(F) = 0.4$
Region	$P^*(R) = 0.1$, $P^*(U) = 0.9$
Car.use	$P^*(C1) = 0.175$, $P^*(C2) = 0.517$, $P^*(C3) = 0.005$, $P^*(C4) = 0.303$

Table 4: Specification of the inputs for South Korean insurance market

After running the algorithm stated in Section 3.3 with the aforementioned input specifications, one can get an adaptive portfolio that comes with the following ratios of target variables summarized in Table 5 and Figure 3. The resulting ratios confirm that the proposed algorithm can effectively replicate specified target variables to mimic the target market accordingly, as we expected.

Variable	Classes	Realized ratio
Insured.age	[16,30) / [30,40) / [40,50) / [50,60) / [60,103]	0.16 : 0.21 : 0.24 : 0.22 : 0.17
Insured.sex	Male / Female	0.59 : 0.41
Region	Rural / Urban	0.1 : 0.9
Car.use	Private / Commute / Farmer / Commercial	0.117 : 0.591 : 0.008 : 0.284

Table 5: Realized ratio of the target variables in the generated portfolio

While the target and original ratios are quite close in the cases of the benchmark variables, one can observe that the distributions of the other variables are not identical to those of the original one, which distinguish the generated portfolio from the original one. Figure 4 shows that the generated and original portfolios have different distributions of two response variables (`NB_Claim`, `AMT_Claim`) where the variability increases as the perturbation input σ increases.

Likewise, Figure 5 presents how the distributions of generated values of telematics variables may differ from those of original values of telematics variables. As shown in Figure 5, it is observed that as the level of perturbation increases, the proportion of random noise in the generated telematics variable also increases. For example, in the case of `Years.noclaims`, its distribution becomes almost uniform when a large value of perturbation parameter is applied.

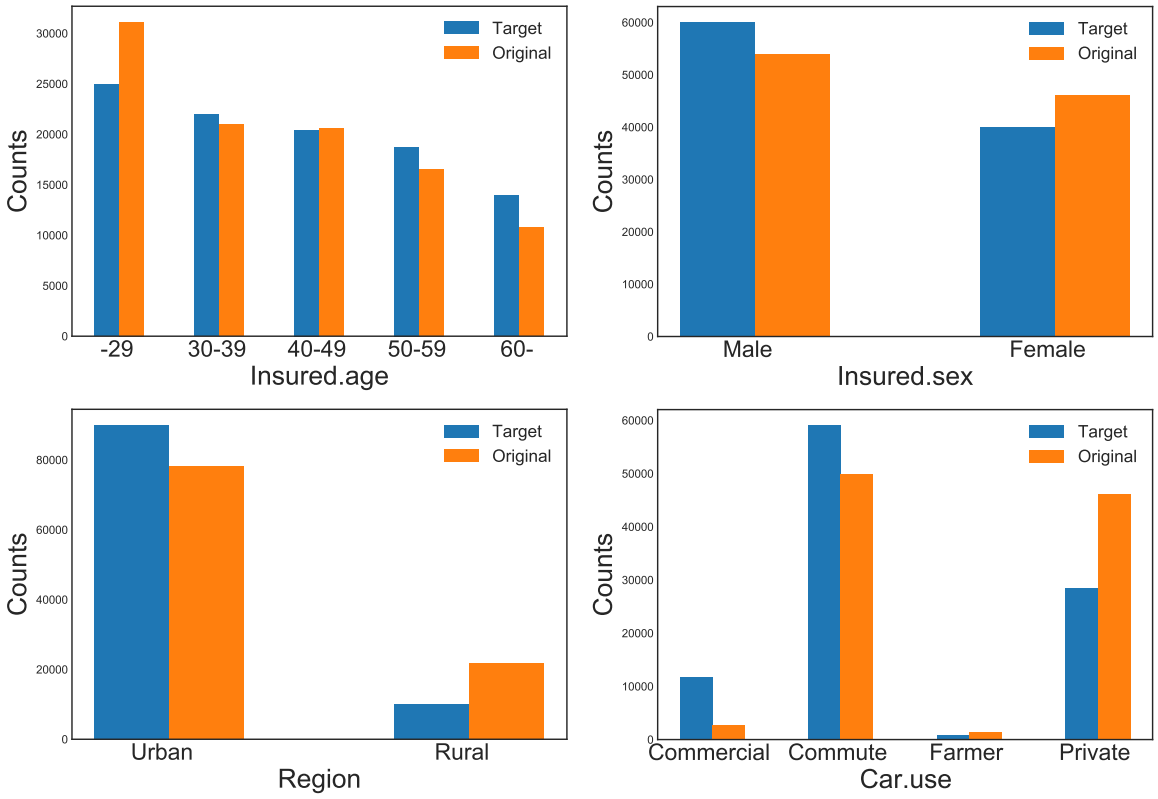


Figure 3: Target ratios (inputs) and original ratios

5 Concluding remarks

In this article, we explored a new algorithm, `SEAT`, which allows to generate an insurance claims data with telematics variables and tailored to match available policy characteristics of the target market. The proposed algorithm utilizes both the ratios of selected policyholder characteristics and colinearity of these variables in the original data as input, which leads to generation of a synthesized claim data that matches to the target market and also is differentiated from the original data. The proposed simulation engine, `SEAT`, is also fully open to public and capable of generating granular dataset with telematics features.

We acknowledge that the proposed algorithm uses the original feature portfolio as the main input so that the behaviors of the synthetic portfolio with the proposed algorithm still heavily depend on the source dataset. Nevertheless, we believe that this research could be meaningful by providing a way to generate telematics claim datasets with flexibility and accessibility, which can encourage both the practitioners and researchers to deepen their understanding and possible uses of telematics data.

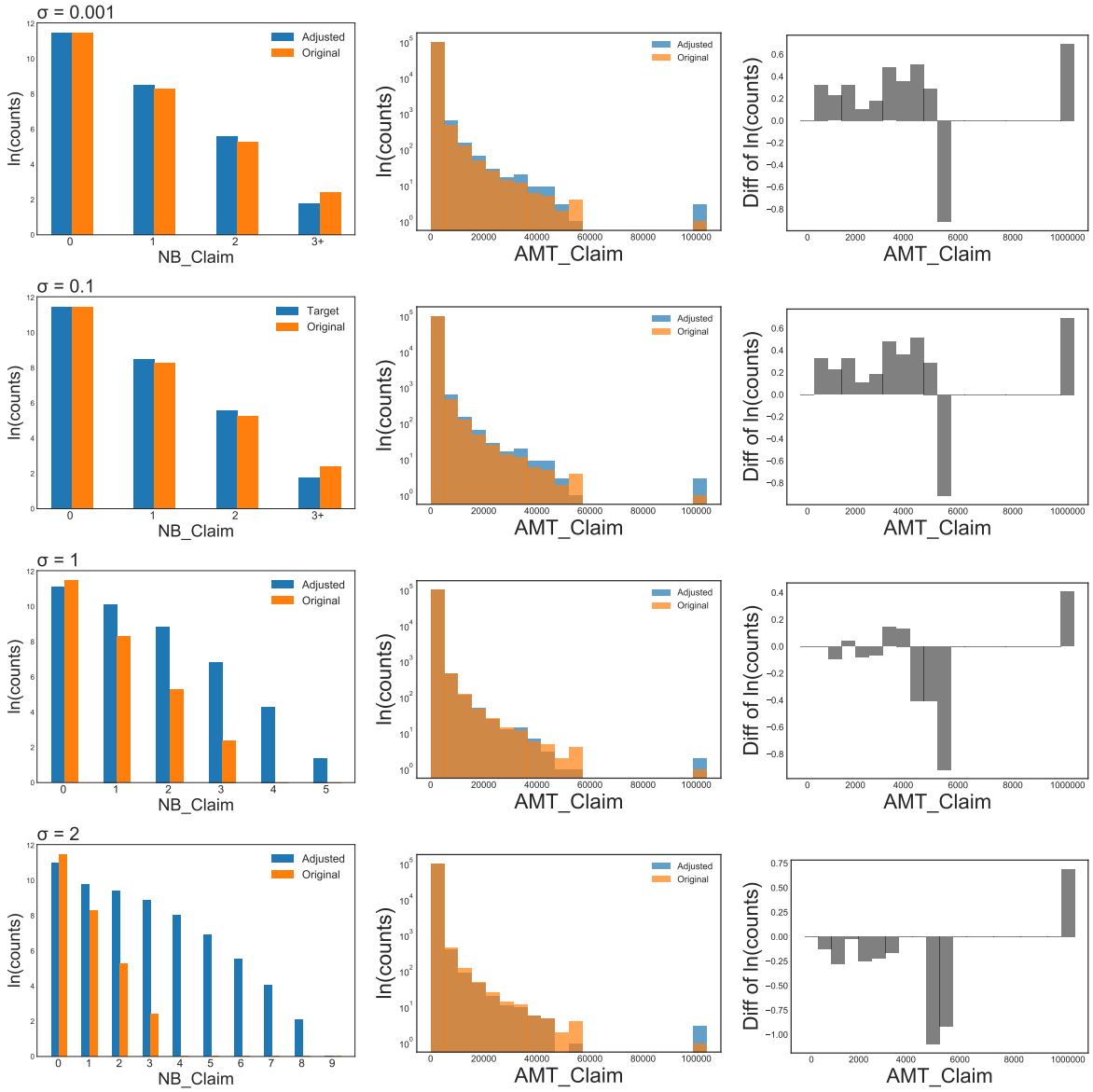


Figure 4: Comparisons of `NB_Claim` (left) and `AMT_Claim` (middle and right) in target and original data

References

Avanzi, B., Taylor, G., Wang, M., and Wong, B. (2021). Synthetic: an individual insurance claim simulator with feature control. *Insurance: Mathematics and Economics*, 100:296–308.

Ayuso, M., Guillen, M., and Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46(3):735–752.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, 73:125–131.

Ayuso, M., Guillen, M., and Pérez-Marín, A. M. (2016). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

Boucher, J.-P., Côté, S., and Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4):54.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cote, M.-P., Hartman, B., Mercier, O., Meyers, J., Cummings, J., and Harmon, E. (2020). Synthesizing property & casualty ratemaking datasets using generative adversarial networks. *arXiv preprint arXiv:2008.06110*.

Duri, S., Elliott, J., Gruteser, M., Liu, X., Moskowitz, P., Perez, R., Singh, M., and Tang, J.-M. (2004). Data protection and data sharing in telematics. *Mobile networks and applications*, 9(6):693–701.

Duri, S., Gruteser, M., Liu, X., Moskowitz, P., Perez, R., Singh, M., and Tang, J.-M. (2002). Framework for security and privacy in automotive telematics. In *Proceedings of the 2nd international workshop on Mobile commerce*, pages 25–32.

Gabrielli, A. and Wüthrich, M. V. (2018). An individual claims history simulation machine. *Risks*, 6(2):29.

Gan, G. and Valdez, E. A. (2018). Nested stochastic valuation of large variable annuity portfolios: Monte carlo simulation and synthetic datasets. *Data*, 3(3):31.

Gao, G., Meng, S., and Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2019(2):143–162.

Gao, G. and Wüthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. *European Actuarial Journal*, 8(2):383–406.

Gerardo, B. D. and Lee, J. (2009). A framework for discovering relevant patterns using aggregation and intelligent data mining agents in telematics systems. *Telematics and Informatics*, 26(4):343–352.

Guillen, M., Nielsen, J. P., Ayuso, M., and Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk analysis*, 39(3):662–672.

Guillen, M., Nielsen, J. P., and Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., and Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, 24(1):141–152.

Han, B.-K. (2016). Discussion about introduction of usage based insurance - focusing on practice guide from 'association of british insurers'. *Korea Insurance Law Journal*, 10:203–247.

Johanson, M., Belenki, S., Jalminger, J., Fant, M., and Gjertz, M. (2014). Big automotive data: Leveraging large volumes of data for knowledge-driven product development. In *2014 IEEE international conference on big data (Big Data)*, pages 736–741. IEEE.

Korean Insurance Research Institute (2021). Korean insurance industry 2021.

Korean Statistical Information Service (2020a). Current status of driver's license holders by gender. https://kosis.kr/statHtml/statHtml.do?orgId=132&tblId=TX_13201_A001&checkFlag=N.

Korean Statistical Information Service (2020b). Current status of driver's license holders by use. https://kosis.kr/statHtml/statHtml.do?orgId=116&tblId=DT_MLTM_1244&checkFlag=N.

Pérez-Marín, A. M., Guillen, M., Alcañiz, M., and Bermúdez, L. (2019). Quantile regression with telematics information to assess the risk of driving above the posted speed limit. *Risks*, 7(3):80.

Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2):70.

So, B., Boucher, J.-P., and Valdez, E. A. (2021). Synthetic dataset generation of driver telematics. *Risks*, 9(4):58.

The Korean Ministry of Land, Infrastructure and Transport (2020). Urbanization rate. https://www.eum.go.kr/web/cp/st/stUpisStatDet.jsp.

The Korean National Police Agency (2020). Current status of driver's license holders by age. https://www.data.go.kr/data/15048419/fileData.do.

Weidner, W., Transchel, F. W., and Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, 6(1):3–24.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1):89–108.

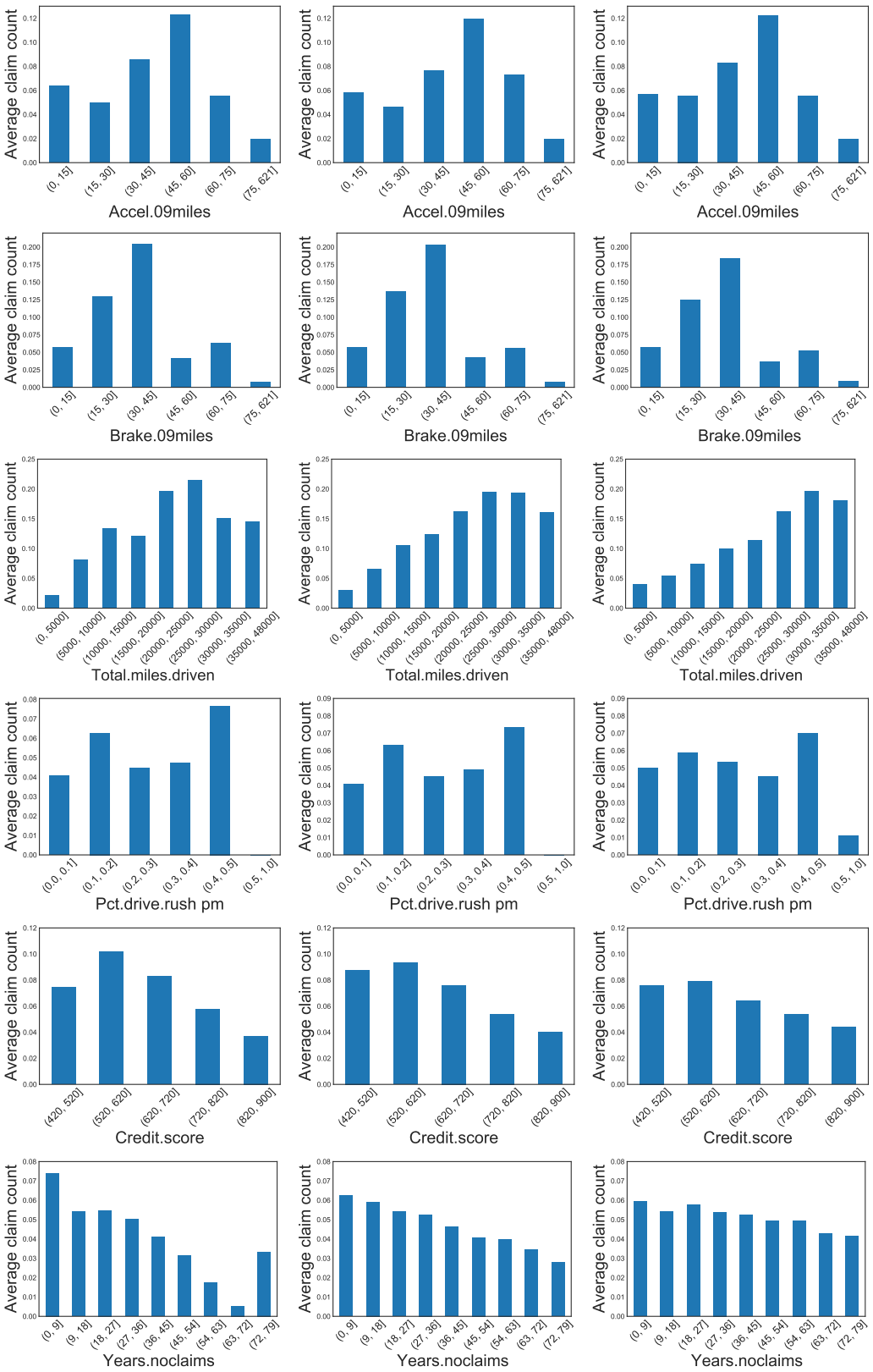


Figure 5: Distributions of generated covariates by varying perturbation parameter: `left`($\sigma = 0.1$), `middle`($\sigma = 1$), `right`($\sigma = 2$)