So, Banghee, and Himchan Jeong. 2025. "Simulation Engine for Adaptive Telematics Data." *Variance* 18 (July). https://doi.org/10.66573/001c.140963.

Financial and Statistical Methods

Simulation Engine for Adaptive Telematics Data

Banghee So[1a], Himchan Jeong[2b]

[1] Department of Mathematics, Towson University, [2] Department of Statistics and Actuarial Science, Simon Fraser University

https://doi.org/10.66573/001c.140963

Variance

Vol. 18, 2025

This article introduces a simulation engine for adaptive telematics (SEAT), which flexibly generates insurance claims datasets from driver telematics information that matches the specific profile of a target market. Generating adaptive telematics data via SEAT is a two-stage process. In the first stage, SEAT uses predetermined distributions of traditional policy characteristics from a target market as inputs and replicates these policy characteristics based on their distributions. In the second stage, SEAT generates the remaining covariates and insurance claims accordingly, based on configurations of the traditional policy characteristics with possible perturbations. We illustrate how SEAT generates an adaptive telematics dataset to match the South Korean insurance market and compare its behavior with the source telematics dataset on which the algorithm is based. We hope that both practitioners and researchers will use this publicly available simulation engine (https://github.com/bheeso/SEAT.git) and adaptive datasets to explore the usefulness of driver telematics data for developing diverse models of usage-based insurance.

Address for Correspondence: himchan_jeong@sfu.ca

1. INTRODUCTION AND MOTIVATION

Recent innovations in technology have made it possible to use telematics devices to track individual automobile vehicle usage data, such as mileage, speed, and acceleration. *Telematics* refers to "the use or study of technology that allows information to be sent over long distances using computers" (*Oxford English Dictionary*). Telematics are widely used in many fields. The insurance industry has been applying telematics to some practices, such as usage-based insurance, which allows them to collect additional information from driving records beyond traditionally observable policy and driver characteristics. Using a plug-in device or mobile application to collect data, insurers can easily gather relevant information from specific policyholders, providing a more sophisticated method for classifying risk.

Inspired by much interest in this practice, telematics use for automobile insurance has been actively studied in the actuarial literature. Ayuso et al. (2014) published one of the earliest studies to address using telematics for pay-as-you-drive insurance. The authors analyzed telematics data and found that vehicle usage differed substantially between novice and experienced young drivers.

Other research has also demonstrated that telematics provide valuable additional data that supplements traditional rating variables. According to Ayuso et al. (2016), gender—a widely used traditional rating variable—is not a significant rating factor after controlling for observed driving factors available through telematics. In a later study, Ayuso et al. (2019) found a significant effect of driving habits on the expected number of claims. Guillen et al. (2019) found that telematics devices help to more precisely predict excess zeros in claim frequencies. Other studies that demonstrated the efficacy of telematics information for improving traditional risk classification models include Boucher (2017), Gao et al. (2019), Pesantez-Narvaez et al. (2019), and Pérez-Marín et al. (2019). Recently, Guillen et al. (2020; 2021) used telematics data to analyze near miss events.

a Dr. Banghee So received her bachelor's and master's degrees in applied statistics from the Yonsei University in Korea and a PhD in actuarial science from the University of Connecticut in 2021. She then joined the Department of Mathematics at Towson University as an assistant professor. She is a Fellow of the Society of Actuaries (FSA). Her research interests include data mining, machine learning, statistical learning, and predictive analytics.

b Himchan Jeong has a PhD in mathematics from the University of Connecticut and is also a Fellow of the Society of Actuaries (SOA). Professionally, Dr. Jeong has authored over 20 peer-reviewed publications, appearing in the well-known actuarial science and statistics journals such as *Insurance: Mathematics and Economics, Journal of Royal Statistical Society: Series A, Scandinavian Actuarial Journal,* and *ASTIN Bulletin.* He has also been awarded grants from actuarial organizations, such as the Canadian Institute of Actuaries and the CAS, as well as the Natural Sciences and Engineering Research Council (NSERC) of Canada. His current research interest is predictive modeling for ratemaking and reserving of property and casualty insurance.

Despite ongoing academic and industry research, access to telematics data has been limited because of privacy concerns; this has made it difficult for actuarial and insurance community members to apply more diverse analytics to telematics data. To address this barrier, So, Boucher, and Valdez (2021) proposed a novel algorithm that generates a synthetic dataset of driver telematics by emulating an existing insurance portfolio. However, the authors' algorithm replicates the distributions of independent and dependent variables almost identically, which might make the use of the resulting synthetic dataset less practical if overall characteristics of the target portfolio differ substantially from the original telematics data. Therefore, we proposed an advanced algorithm of adaptive telematics data generation that incorporates prior knowledge of an insurance market's or company's specific characteristics. We used the synthetic dataset from So, Boucher, and Valdez (2021) as the source data for our algorithm and generated an output dataset with modified distributions of the observed responses and covariates. This approach could provide a concrete method for generating synthetic telematics datasets based on insurers' various needs.

The remaining sections are organized as follows. Section 2 introduces the general concept and characteristics of telematics information and synthetic datasets. Section 3 covers our main work, which includes constructing the simulation engine for adaptive telematics (SEAT) data, along with relevant theoretical foundations and actual implementation. Section 4 illustrates an empirical application of the proposed method by generating a portfolio that matches the South Korean insurance market. We also discuss the characteristics of the generated portfolio compared with the original portfolio. We conclude with remarks in Section 5.

2. TELEMATICS AND SYNTHETIC DATA

As mentioned in Section 1, vehicle usage telematics data are collected and transmitted on a real-time basis; therefore, telematics data collection and processing differs substantially from that of traditional data. According to Johanson et al. (2014), a driving signal dataset covering 1,000 vehicles will generate about 560 gigabytes per day. Since a typical automobile insurance portfolio includes millions of drivers, problems may arise regarding the effectiveness of data recording, data privacy (Duri et al. 2002, 2004), and feature extraction. Telematics data feature extraction is illustrated in Figure 1.

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

Recently, Wüthrich (2017) and Gao and Wüthrich (2018) suggested a different framework for telematics data feature extraction and dimension reduction. Since telematics data inherently form a continuum of observations, the authors considered the velocity-acceleration heatmaps and analyzed them via a k-means clustering algorithm to classify

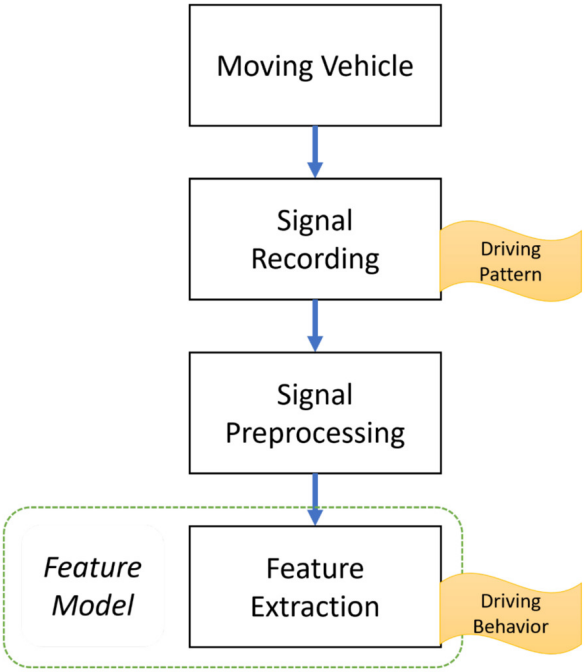

Figure 1. Feature extraction from telematics data (Weidner, Transchel, and Weidner 2016).

car drivers' risk. Henceforth, we focus on emulation and analysis of a summarized form of telematics data rather than of raw data, as found in most actuarial studies.

However, finding a publicly accessible telematics dataset, even in a summarized form, has been extremely difficult, owing to privacy concerns of insurers who are reluctant to provide their driving records to the public. Given this, insurers are increasingly interested in using synthetic data, because it is free from privacy issues yet maintains the original dataset's essential characteristics. Synthetic data can also be easily implemented for construction and validation of statistical models that improve risk classifications. For example, Gan and Valdez (2018) created a synthetic dataset of a large variable annuity portfolio that can be used to develop annuity valuation or hedging techniques such as metamodeling. Gabrielli and Wüthrich (2018) proposed an individual claims history simulation machine that helps researchers calibrate their own individual or aggregate reserving models. Cote et al. (2020) applied generative adversarial networks to synthesize a property and casualty ratemaking dataset, which could be used for predictive analytics in the ratemaking process. Avanzi et al. (2021) introduced an individual insurance claims simulator with feature control, which allows modelers to assess the validity of their reserving methods by back-testing.

To the best of our knowledge, So, Boucher, and Valdez (2021) is the first study to address synthesizing a dataset that includes features engineered from an actual telematics dataset. The authors used a three-stage process that applies various machine learning algorithms. First, they generated feature variables of a synthetic portfolio via an extended version of the synthetic minority oversampling technique (Chawla et al. 2002). Second, they simulated a correspond-

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

ing number of claims via binary classifications using feedforward neural networks. Finally, they simulated corresponding aggregated amounts of claims via regression feedforward neural networks.

Despite its novelty, the extended synthetic minority oversampling algorithm for feature generation produced almost identical distributions of the traditional and telematics feature variables from the synthetic and original datasets, as shown in the Appendix of So, Boucher, and Valdez (2021). This poses a concern for researchers and practitioners wishing to use a synthetic dataset, given disparities in target market and original data feature portfolios. Therefore, we propose an innovative simulation engine, SEAT, that uses target market prior information to synthesize a feature portfolio similar to that of the target market.

3. SIMULATION OF ADAPTIVE TELEMATICS DATA

3.1. DESIRABLE CHARACTERISTICS OF A SYNTHETIC TELEMATICS DATASET

Before introducing SEAT in detail, we will discuss the agreed-upon desirable characteristics of a synthetic dataset. First, the dataset should be accessible to the public. As previously mentioned, limited access to telematics data is a major obstacle to developing and back-testing ratemaking methods with telematics features. Second, the dataset should be flexible enough to satisfy the needs of modelers with diverse and specific interests. Finally, dataset granularity should be assured so that the data can be used to train, test, and apply various features related to individual risk classification via predictive modeling.

Therefore, we aimed to develop a publicly available method that allows anyone to access the sample dataset, data generation routine, and source codes from the following link (https://github.com/bheeso/SEAT.git). The resulting dataset is fully granular and flexible in that it can match various target market profiles of interest. In future research, we hope to incorporate other important characteristics, such as longitudinality and multiple lines of business to emulate serial and/or between-coverage dependence.

3.2. DESCRIPTION OF THE SOURCE DATA

As discussed earlier, SEAT uses synthetic insurance claims data from So, Boucher, and Valdez (2021; http://www2.math.uconn.edu/~valdez/data.html, accessed on April 30, 2025) as the data source, which consists of traditional and telematics features and two response variables.

The source dataset contains 52 variables, which can be categorized into three groups:

1. 11 traditional features, such as insurance exposure, age of driver, and main use of the vehicle.
2. 39 telematics features, such as total distance driven and number of sudden accelerations.
3. Two response variables that describe the claim frequency and aggregated claim amounts.

Table 2 shows the name, description, and data attributes of the available features in the dataset. Note that feature attributes are important because they affect the data generation scheme that involves random perturbation, as elaborated in Step 4 of Section 3.3. For detailed information and preliminary analysis of the dataset, see Section 3 of So, Boucher, and Valdez (2021).

3.3. ADAPTIVE DATA GENERATION SCHEME

The proposed algorithm in SEAT uses predetermined distributions of the traditional covariates such as sex, region, age, and main use of the insured vehicle, which are easily accessible in the target market. In our source dataset, they are named `Insured.sex`, `Region`, `Insured.age`, and `Car.use`, respectively.

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 to the algorithm. Table 3 shows how these inputs are defined. For example, $P^*(A1)$ (in Table 3) indicates the ratio of insured who are between ages 16 and 30. We produced an adaptive portfolio by applying the SEAT algorithm, based on the ratios of the four variables.

When the input values are provided, the following algorithm generates data that are adapted from the source data:

- Step 1: Calculate conditional distributions of the benchmark covariates (see Figure 2) using source data and inputs. The four variables may be colinear, so we used modified conditional distributions to reflect such collinearities in the algorithm. To obtain the modified conditional distributions, we used the following probability rules where $P^*(\cdot)$ represents the probabilities for the target distribution. (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.)

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

Table 2. Source dataset variable names and descriptions.

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 accelerations 6/8/9/.../14 mph/s per 1,000 miles	Ordinal
	`Brake.xxmiles`	Number of sudden brakes 6/8/9/.../14 mph/s per 1,000 miles	Ordinal
	`Left.turn.intensityxx`	Number of left turns per 1,000 miles with intensity 08/09/10/11/12	Ordinal
	`Right.turn.intensityxx`	Number of right turns per 1,000 miles 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 3. Description of inputs associated with the benchmark covariates.

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	$P^*(C1), P^*(C2), P^*(C3), P^*(C4)$

and

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

where

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

and $P(\cdot)$ is 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 $\frac{P^*(U)}{P(U)}$ adjustment factors since they play an important role in adjusting ratios of the original portfolio to ratios in the adaptive portfolio. 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`. First, one can sample a random observation for `Insured.sex` from its benchmark distribution [$P^*(\texttt{Insured.sex} = F)$ and $P^*(\texttt{Insured.sex} = M)$]. If a uniform random number is categorized into female (`F`), then, in the next stage, a newly generated uniform random number is categorized based on the conditional distribution of `Insured.sex` [$P^*(\texttt{Region} = R \mid \texttt{Insured.sex} = F)$ and $P^*(\texttt{Region} = U \mid \texttt{Insured.sex} = F)$].

- Step 3: By repeating Step 2 N times, obtain 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 with Gaussian white noise, conditional on each configuration of the benchmark covariates. Note that the number of possible combinations of the benchmark covariates increases exponentially, depending on the number of covariates and levels of a covariate. Therefore, many cases may have configurations with sparse empirical observations. To avoid repeatedly producing the same samples, we twisted samples slightly using Gaussian white noise. For example, for configuration $\mathcal{C}_i = \{\texttt{Car.use} = F, \texttt{Region} = R, \texttt{Insured.age} = A1, \texttt{Insured.sex} = C1\}$, the corresponding independent variables (both telematics and remaining traditional features) $\mathcal{T}_i = (T_i^{(1)}, \ldots, T_i^{(p_\mathcal{T})})$, number of claims N_i, and total amount 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 and remaining traditional features, $\tilde{F}_X(\cdot|\mathcal{C})$ is the empirical distribution of feature X given the configuration of the benchmark covariate $\mathcal{C}$, and σ is the input that controls the degree of random perturbation. For simplicity and convenience of employing the SEAT algorithm, we use the same σ to all covariates. However, to give an equivalent degree of randomness to all covariates, their scales are adjusted by centering and scaling with the median and interquartile range. For example, for a variable X, the scale-adjusted variable is defined as $\frac{X - Q_2(X)}{Q_3(X) - Q_1(X)}$ where $Q_1(X), Q_2(X)$, and $Q_3(X)$ denote the first, second, and third quartile of the observed values of X, respectively. Note that for 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 number, respectively. We also set an upper limit and lower limit of each variable so that values are inside the expected interval.

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

4. EMPIRICAL APPLICATION: SYNTHETIZED TELEMATICS DATA FOR THE SOUTH KOREAN INSURANCE MARKET

To test the effectiveness of SEAT, we generated a synthetized telematics dataset tailored to the South Korean insurance market with predetermined inputs and analyzed its properties compared with the original telematics dataset. We used the South Korean insurance market as a case study, but the proposed algorithm is applicable to any national or regional market, as well as to any industry or company, as it can be adjusted based on the specific interests of the modeler.

The South Korean insurance market has grown rapidly in recent years, ranking seventh in total premium volume in 2020 (Korean Insurance Research Institute 2021), with a global market share of 3.1%. Nevertheless, the use of telematics data both in actuarial practice and in research is still developing in South Korea. Although Han (2016) discussed regulatory and legal issues regarding telematics data for usage-based insurance in the South Korean insurance market, follow-up research on implementation of ratemaking methods with telematics data is lacking, partially due to the scarcity of publicly available data. Furthermore, only one company, Carrot Insurance, actively uses driver telematics information in their ratemaking scheme. We therefore extracted basic profiles of the South Korean insurance market from various sources (The Korean National Police Agency 2020; Korean Statistical Information Service 2020a, 2020b; The Korean Ministry of Land, Infrastructure and Transport 2020) and used these as the inputs to generate the adaptive portfolio. Table 4 provides the input specification of the benchmark covariates used in dataset generation.

Running the algorithm (as described in Section 3.3) with the input specifications produced an adaptive portfolio; the resulting ratios of target variables are summarized in Table 5 and Figure 3. The ratios confirm that the proposed algorithm effectively replicated specified benchmark covariates to mimic the target market accordingly, as we expected.

While the target and original ratios are quite close in the benchmark covariates, the other variable distributions are not identical to those of the originals, which distinguishes the generated portfolio from the original. Figure 4 shows that the distributions of the two response variables differ between the generated and original portfolios (NB_Claim, AMT_Claim), where the variability increases as the perturbation input σ increases.

Similarly, Figure 4 shows how the distributions of the covariates' generated values may differ from those of the covariates' original values. As the level of perturbation increases, the influence of random noise in the generated covariates also increases. For example, in the case of Years.noclaims, the distribution becomes almost uniform when a large perturbation parameter is applied.

5. CONCLUDING REMARKS

This article explored a new algorithm, SEAT, which allows users to generate insurance claims datasets with telematics

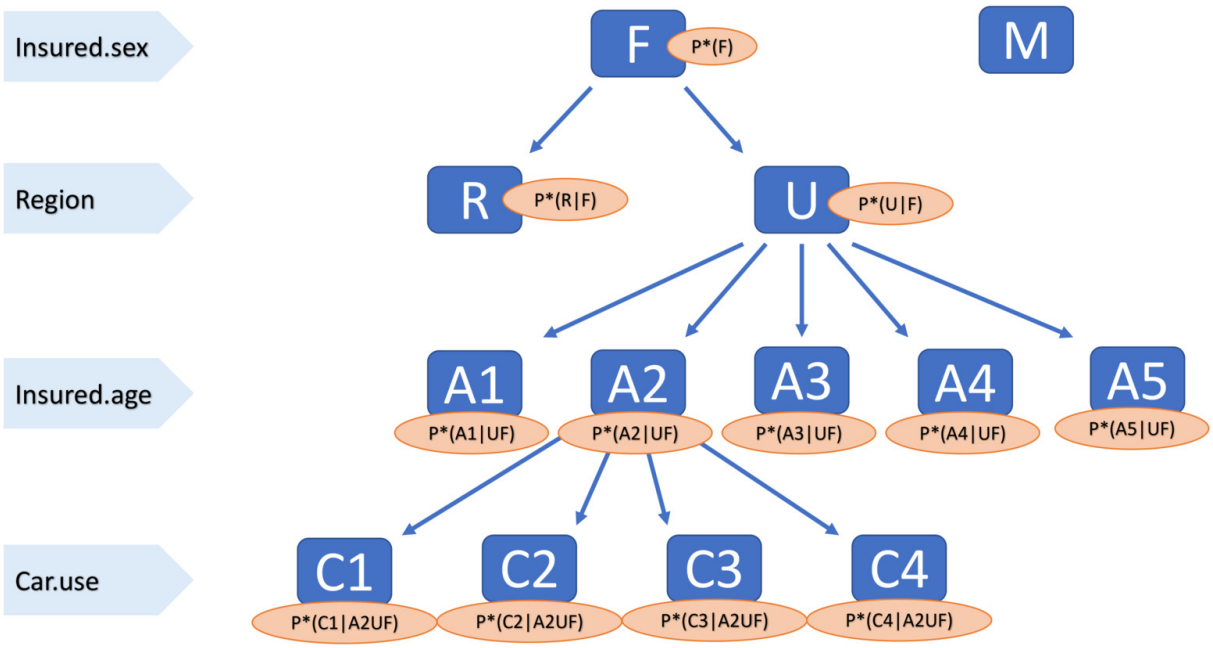

Figure 2. A branch of generating conditional distributions with the benchmark covariates.

Algorithm 1. SEAT.

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

Output: an adaptive dataset matching the target feature portfolio (size is N)

Import the data introduced in So et al. (2021) as the source dataset;
Calculated conditional distributions, $P^*(R \mid F), P^*(U \mid F), P^*(R \mid M), P^*(U \mid M), P^*(A1 \mid RF)$, $P^*(A2 \mid RF), \cdots, P^*(A5 \mid UM), P^*(C1 \mid A1RF), \cdots, P^*(C4 \mid A5UM)$;

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 the source data, randomly sample an observation having $\mathcal{C}_i$ with white Gaussian noise (σ) as follows:

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

end
Return the adaptive dataset: $\{(\mathcal{C}_1, \mathcal{T}_1, N_1, S_1), (\mathcal{C}_2, \mathcal{T}_2, N_2, S_2), \cdots, (\mathcal{C}_N, \mathcal{T}_N, N_N, S_N)\}$.

features that are tailored to match the policy characteristics of a target market. The proposed algorithm employs both the ratios of selected policyholder characteristics and collinearity of these benchmark covariates in the original data as inputs. This generates a synthetized claims dataset that matches the target market where the resulting dataset is not identical to the original. The proposed simulation en-

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

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 5. Realized ratio of the benchmark covariates in the generated portfolio.

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

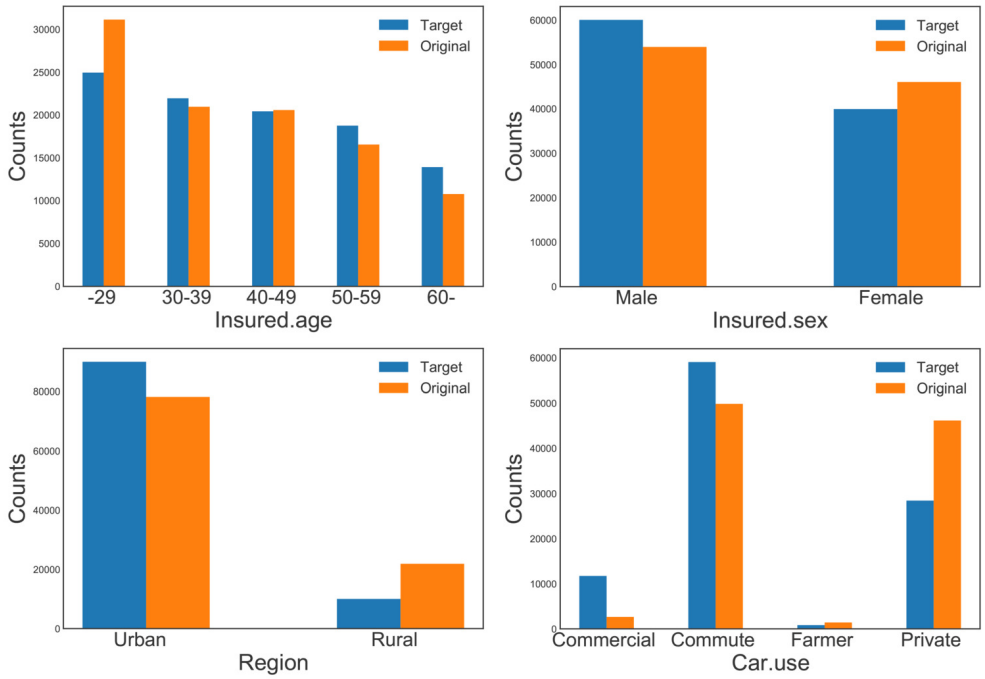

Figure 3. Target ratios (inputs) and original ratios.

gine, SEAT, is also fully accessible to the public and can generate granular datasets with telematics features.

We acknowledge that the proposed algorithm uses the original feature portfolio as the main source; therefore the behavior of the synthetic portfolio is heavily dependent on the source dataset. Nevertheless, this research is meaningful in that it provides a novel method of producing telematics claims datasets with flexibility and accessibility, which encourages both practitioners and researchers to deepen their understanding of telematics data, and provides insight into the potential uses of existing telematics data. Finally, the procedure described in this paper could be expanded to many different types of data in addition to telematics. For instance, it could be applied as an extended version of the Casualty Actuarial Society (2018) CAS loss simulator to simulate data for a claims-level loss reserving model with individual policy characteristics.

ACKNOWLEDGMENTS

The authors thank Daesan Shin Yong Ho Memorial Society for financial support through its Insurance Research Grant program.

Submitted: June 15, 2022 EDT. Accepted: November 30, 2022 EDT. Published: July 10, 2025 EDT.

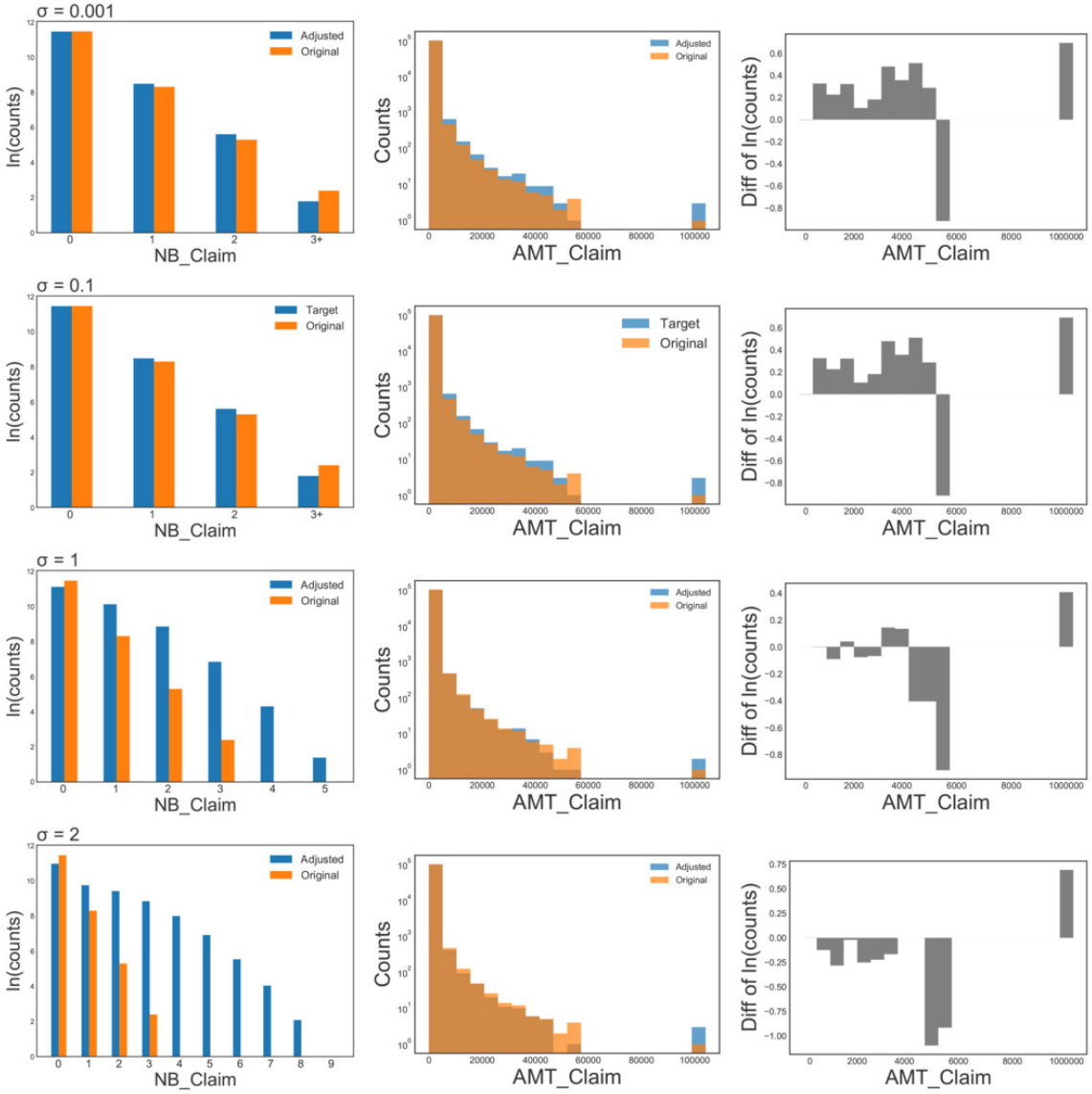

Figure 4. Comparisons of `NB_Claim` (left) and `AMT_Claim` (middle and right) in target and original data.

REFERENCES

Avanzi, Benjamin, Greg Taylor, Melantha Wang, and Bernard Wong. 2021. "SynthETIC: An Individual Insurance Claim Simulator with Feature Control." *Insurance: Mathematics and Economics* 100:296–308.

Ayuso, Mercedes, Montserrat Guillen, and Jens Perch Nielsen. 2019. "Improving Automobile Insurance Ratemaking Using Telematics: Incorporating Mileage and Driver Behaviour Data." *Transportation* 46 (3): 735–52.

Ayuso, Mercedes, Montserrat Guillén, and Ana María Pérez-Marín. 2014. "Time and Distance to First Accident and Driving Patterns of Young Drivers with Pay-as-You-Drive Insurance." *Accident Analysis & Prevention* 73:125–31.

Ayuso, Mercedes, Montserrat Guillen, and Ana María Pérez-Marín. 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, Jean-Philippe, Steven Côté, and Montserrat Guillen. 2017. "Exposure as Duration and Distance in Telematics Motor Insurance Using Generalized Additive Models." *Risks* 5 (4): 54.

Casualty Actuarial Society. 2018. "CAS Loss Simulator 2.0." https://www.casact.org/publications-research/research/loss-simulation-model-and-documentation.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16:321–57.

Cote, Marie-Pier, Brian Hartman, Olivier Mercier, Joshua Meyers, Jared Cummings, and Elijah Harmon. 2020. "Synthesizing Property & Casualty Ratemaking Datasets Using Generative Adversarial Networks." *arXiv Preprint arXiv:2008.06110*.

Duri, Sastry, Jeffrey Elliott, Marco Gruteser, Xuan Liu, Paul Moskowitz, Ronald Perez, Moninder Singh, and Jung-Mu Tang. 2004. "Data Protection and Data Sharing in Telematics." *Mobile Networks and Applications* 9 (6): 693–701.

Duri, Sastry, Marco Gruteser, Xuan Liu, Paul Moskowitz, Ronald Perez, Moninder Singh, and Jung-Mu Tang. 2002. "Framework for Security and Privacy in Automotive Telematics." In *Proceedings of the 2nd International Workshop on Mobile Commerce*, 25–32.

Gan, Guojun, and Emiliano A. Valdez. 2018. "Nested Stochastic Valuation of Large Variable Annuity Portfolios: Monte Carlo Simulation and Synthetic Datasets." *Data* 3 (3): 31.

Gao, Guangyuan, Shengwang Meng, and Mario V. Wüthrich. 2019. "Claims Frequency Modeling Using Telematics Car Driving Data." *Scandinavian Actuarial Journal* 2019 (2): 143–62.

Gao, Guangyuan, and Mario V. Wüthrich. 2018. "Feature Extraction from Telematics Car Driving Heatmaps." *European Actuarial Journal* 8 (2): 383–406.

Gerardo, Bobby D., and Jaewan Lee. 2009. "A Framework for Discovering Relevant Patterns Using Aggregation and Intelligent Data Mining Agents in Telematics Systems." *Telematics and Informatics* 26 (4): 343–52.

Guillen, Montserrat, Jens Perch Nielsen, Mercedes Ayuso, and Ana M. Pérez-Marín. 2019. "The Use of Telematics Devices to Improve Automobile Insurance Rates." *Risk Analysis* 39 (3): 662–72.

Guillen, Montserrat, Jens Perch Nielsen, and Ana M. Pérez-Marín. 2021. "Near-Miss Telematics in Motor Insurance." *Journal of Risk and Insurance* 88:569–89.

Guillen, Montserrat, Jens Perch Nielsen, Ana M. Pérez-Marín, and Valandis Elpidorou. 2020. "Can Automobile Insurance Telematics Predict the Risk of Near-Miss Events?" *North American Actuarial Journal* 24 (1): 141–52.

Han, Byung-Kyu. 2016. "Discussion about Introduction of Usage Based Insurance - Focusing on Practice Guide from 'Association of British Insurers.'" *Korea Insurance Law Journal* 10:203–47.

Johanson, Mathias, Stanislav Belenki, Jonas Jalminger, Magnus Fant, and Mats Gjertz. 2014. "Big Automotive Data: Leveraging Large Volumes of Data for Knowledge-Driven Product Development." In *2014 IEEE International Conference on Big Data (Big Data)*, 736–41. IEEE.

Korean Insurance Research Institute. 2021. "Korean Insurance Industry 2021." Korean Insurance Research Institute.

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.

———. 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, Ana M., Montserrat Guillen, Manuela Alcañiz, and Lluís Bermúdez. 2019. “Quantile Regression with Telematics Information to Assess the Risk of Driving above the Posted Speed Limit.” *Risks* 7 (3): 80.

Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. “Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression.” *Risks* 7 (2): 70.

So, Banghee, Jean-Philippe Boucher, and Emiliano A. Valdez. 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, Wiltrud, Fabian W. G. Transchel, and Robert Weidner. 2016. “Classification of Scale-Sensitive Telematic Observables for Riskindividual Pricing.” *European Actuarial Journal* 6 (1): 3–24.

Wüthrich, Mario V. 2017. “Covariate Selection from Telematics Car Driving Data.” *European Actuarial Journal* 7 (1): 89–108.