UDC 004.852

ANALYSIS OF DATA UNCERTAINTIES IN MODELING AND FORECASTING OF ACTUARIAL PROCESSES

Panibratov R. S. – Postgraduate student of the Institute for Applied System Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

ABSTRACT

Context. Analysis of data uncertainties in modeling and forecasting of actuarial processes is very important issue because it allows actuaries to efficiently construct mathematical models and minimize insurance risks considering different situations.

Objective. The goal of the following research is to develop an approach that allows for predicting future insurance payments with prior minimization of possible statistical data uncertainty.

Method. The proposed method allows for the implementation of algorithms for estimating the parameters of generalized linear models with the preliminary application to data of the optimal Kalman filter. The results demonstrated better forecast results and more adequate model structures. This approach was applied successfully to the simulation procedure of insurance data. For generating insurance dataset the next features of clients were used: age; sex; body mass index (applying normal distribution); number of children (between 0 and 5); smoker status; region (north, east, south, west, center); charges. For creating the last feature normal distribution with known variance and a logarithmic function, exponential distribution with the identity link function and Pareto distribution with a known scale parameter and a negative linear function were used.

Results. The proposed approach was implemented in the form of information processing system for solving the problem of predicting insurance payments based on insurance data and with taking into account the noise of the data.

Conclusions. The conducted experiments confirmed that the proposed approach allows for more adequate model constructing and accurate forecasting of insurance payments, which is important point in the analysis of actuarial risks. The prospects for further research may include the use of this approach proposed in other fields of insurance related to availability of actuarial risk. A specialized intellectual decision support system should be designed and implemented to solve the problem by using actual insurance data from real world in online mode as well as modern information technologies and intellectual data analysis.

KEYWORDS: actuarial risk, generalized linear models, optimal Kalman filter, exponential family of distributions, simulation, iterative-recursive weighted least squares method, Adam method, Monte Carlo for Markov chains.

ABBREVIATIONS

NPV is a net present value;
MEBN is a multi-entity Bayesian networks;
EVT is an extreme value theory;
GLM is a generalized linear models;
GLMC is a generalized linear models with credibility;
MCMC is a Markov chain Monte Carlo method;
IRWLS is an iterative-recursive weighted least squares.
MSE is a mean squared error.
RMSE is a root mean squared error.
MAE is a mean absolute error,
Adam is an adaptive moment estimation.

NOMENCLATURE

A is a system dynamic matrix;

$x(n)$ is a vector of states at time step $n > 0$;

$\tilde{x}(n)$ is a vector estimate of states at time step $n > 0$;

B is a matrix of control coefficients;

$u(n)$ is a vector of controls at time step $n > 0$;

$w(n)$ is a noise vector at time step $n > 0$, which has a normal distribution with mean vector with all zero values and covariance matrix Q;

Q is a covariance matrix of state disturbances;

$z(n)$ is a vector of measurements of output variables at time step $n > 0$;

H is a matrix of observation coefficients;

$v(n)$ is a vector of random measurement noise values at time step $n > 0$, which has a normal distribution with mean zero vector and covariance matrix R;

R is a matrix of measurement errors;

$P(n)$ is a covariance matrix of errors of state vector estimates at time step $n > 0$;

$K(n)$ is a filter's matrix optimum coefficient at time step $n > 0$;

I is an identity matrix;

$a(\bullet), b(\bullet), c(\bullet,\bullet)$ are functions that are defined at the outset in exponential family of distributions;

θ is a parameter associated with mean values;

φ is a scale parameter associated with variance;

y is a target variable for insurance charges and set of financial processes;

η is a linear predictor;

X is a matrix of covariates;

β is a estimated parameter of GLM;

g is a link function

E is an expected value;

x_m is a scale parameter for Pareto distribution;

σ is a standard deviation for normal distribution.

INTRODUCTION

The existence of factors that prevent possibilities from having deterministic outcomes is implied by uncertainty,

DOI 10.15588/1607-3274-2024-2-5

and it is unknown to what extent these factors may have an impact on the outcomes.

Either a practical or abstract theoretical study of the circumstances for the presence of uncertainty can be carried out, depending on the decision-making perspective that is applied to a particular case. For instance, several mathematical models are employed at the abstract theoretical level, while an evaluation of the quantity of information needed for decision selection is done at the practical level. Selection of these models considers the likelihood of their development in particular scenarios. Information entropy may be used to estimate the quantity of information needed to characterize the uncertainty of the selection scenario.

The uncertainty category is defined by few variable characteristics that characterize many kinds of uncertainties, such as situational, political, social in nature global, and so on. Determining the degree of analysis and the kinds of uncertainties being taken into account is essential to solving the challenges associated with decision-making in the face of uncertainty.

It should be highlighted that uncertainty is frequently limited to the absence of comprehensive knowledge about a particular object. Indeed, uncertainty is not limited to inadequate understanding about object states. In addition, it is occasionally feasible to take into account the ambiguity of the decision-selection criteria and the objectives.

The amount of alternative possibilities and the variety and quantity of criteria used to evaluate these options define the degree of decision-making complexity in many real-world situations.

Since genuine risks and uncertainty are a part of the past, present, and future of analyzed process development, they must be considered in all actions that have an impact on the goals of the organization. Risk and uncertainty are present in all economic activity in varying amounts, but no matter how thorough the risk management, uncertainty cannot be totally removed. Unexpected circumstances and interdependencies might arise at any time. Such unanticipated occurrences may result in deviations that radically alter the data arrangement. As a result, uncertainty can become a risk factor when it results from incomplete information or from using sources that are frequently at odds with the real circumstances of a company or the competitive market [1].

It should be highlighted that a variety of uncertainties, taken together to produce a specific complex of uncertainties known as systemic uncertainty, are frequently present in actual practical situations involving decision-making.

The object of study is the process of search the best approach which allows to analyze actuarial risk more efficiently. It is proposed to generate insurance indicators and target variables randomly with adding noise to simulate real-world data, because they are not always publicly available. Therefore, it is proposed to implement approach, which allows to forecast insurance indicators more efficiently by reducing uncertainty.

The subject of study are methods for forecasting insurance data.

The purpose of the work is to implement approach, which allows to reduce uncertainty during the solving task of forecasting the insurance indicators.

1 PROBLEM STATEMENT

For the class of financial processes $\{y(\bullet)\}$ with a generalized form of the probability distribution:

$$f(y,\theta,\varphi)=\exp\{\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi),$$

where $a(\bullet),b(\bullet),c(\bullet,\bullet)$ are functions, that correspond to a certain distribution law; y is a dependent variable; θ is a canonical parameter or function of some parameter of a certain distribution; φ is the variance parameter. The following distribution laws are allowed: normal, Poisson, binomial, inverse Gaussian, gamma, exponential.

The function $b(\bullet)$ assumes special significance in generalized linear models, because it describes the relationship between the mean value of μ_y and the variance of the process $\{y\}:\sigma_y^2$. If φ is known, then it is an exponential model with the canonical parameter θ. Also, the exponential distribution can be two-parameter, if φ is unknown.

It is proposed to determine and minimize the impact of statistical data regarding the dependent variable $\{y\}$, which lead to deterioration of the results of estimation of the structure and parameters of mathematical models and estimates of forecast calculated on the basis of these models. In this case, the following types of models $\{M_i\}, i=1,2,3,...$ are possible: linear regression, variance and covariance analysis, Log-linear models for the analysis of random tables, probit/logit models, Poisson regression.

2 REVIEW OF THE LITERATURE

As of right now, there are no widely applicable methods for accounting for the uncertainties of the majority of types that are now in use that can be effectively used to solve the aforementioned real-world issues. Generally speaking, the current techniques for handling uncertainties allow for the consideration of certain particular kinds of uncertainties to enhance the quality of the outcome. Thus, for instance, optimal Kalman filter allows for optimum estimates of the process's state to be obtained against the backdrop of negative random effects by accounting for and minimizing the influence of state disturbances and measurement noise.

The author of [2] provided evidence that determining the measurement uncertainty of every approved analytical test process should be viewed as a valuable completion that adds value rather than as an extra burden. Evaluation

DOI 10.15588/1607-3274-2024-2-5

and comparison of a result with other outcomes are made possible by measurement uncertainty. Communicating the positive meaning of "measurement uncertainty" to clients and the head of authority is crucial. Declaring an excessive amount of uncertainty based just on conjecture is illogical.

Data mining algorithms such as neural networks, evolutionary methods, informed search and space exploration targeted at solving optimization problems, mathematical logic, decision trees, and some others are very helpful in the fighting against uncertainties. With the option to include expert estimates, Bayesian networks – probabilistic models in the form of directed acyclic graphs with the variables of the processes under investigation at their vertices–are an incredibly powerful tool for data analysis.

Multi-Entity Bayesian Networks (MEBN) were presented by the authors in [3]. Given any consistent collection of finitely many first-order phrases, its logic may assign a conditional probability distribution and, conversely, assign probabilities to any set of sentences in first-order logic in a logically coherent manner. That is, MEBN logic can assign a probability to everything that can be represented in first-order reasoning. It is not easy to obtain complete first-order expressive capability in the Bayesian logic. Representing an unlimited or potentially infinite number of random variables is necessary for this, some of which could have an unbounded or potentially infinite number of possible values. Furthermore, we must be able to express random variables with potentially infinite or unbounded parents as well as recursive definitions. More challenges arise from possible random variables that take values in uncountable sets, like the real numbers.

According to MEBN logic, the environment is made up of entities with qualities and relationships to other entities. The features of entities and the connections between them are represented by random variables. MEBN theories are collections of MEBN fragments arranged to represent knowledge about qualities and connections. Given their parents in the fragment network and the context nodes, a MEBN fragment provides the conditional probability distribution for instances of its resident random variables. Any collection of MEBN pieces that together satisfy consistency conditions guaranteeing the presence of a distinct joint probability distribution over instances of the random variables each MEBN fragment in the collection represents is referred to as a MEBN theory [3].

In [4], the assessment of risky investment choices is predicted on techniques that have developed over time to account for both project risk and flexibility. The first steps in project evaluation were calculating the project's net present value (NPV) using the proper discount rate. In recent times, managers have been able to ascertain the proper modifications in project value estimations that represent flexibility, or the chance to respond to unforeseen circumstances and surprises, thanks to the instruments of decision trees and actual alternatives. These techniques offer a sophisticated manner of appraising the value of this flexibility.

The authors of [5] examined the reliability and precision of forecasts in a wide range of topics in the scientific and social sciences. Because they are subject to human biases and limitations, judgmental predictions are no more reliable than statistical ones. As long as forecasts are independent and gathered from a variety of sources, combining them appears to increase accuracy. This is especially true for judgmental forecasts, where averaging of several forecasts typically yields forecasts that are more accurate than the best individual forecasts while simultaneously lowering the variation of predicting errors. On the other hand, both statistical models and subjective forecasters often grossly underestimate uncertainty. The authors outlined two main categories of forecasting scenarios that call for various methodologies and models. Predicting normal conditions in a steady, stable context with known patterns and linkages is referred to as the first. The second occurs in peculiar circumstances with ephemeral, shifting patterns. It should be underlined that booms in business and economic recessions and crises cannot be viewed as anomalies; instead, they need to be forecasted using a different acceptable methodology and adequate model [6].

Explorers face many challenges and issues as a result of the significant variation in prediction accuracy and uncertainty over different time horizons. Additionally, the degree of ambiguity and precision differs across different fields. Normal-condition forecasting errors are thin-tailed, but unusual-condition forecasting errors exhibit radically different behavior, with fat tails. Extreme Value Theory (EVT) has shown to be a useful tool for scientists in estimating uncertainty and producing realistic risk assessments that account for fat tail errors while avoiding the pitfall of average assessments, which drastically underestimate risk and uncertainty. Their results have a lot of promise today and in the future and can be used in different forecasting contexts [5].

3 MATERIALS AND METHODS

In particular, the adaptive Kalman filter is a pretty useful tool for assessing and accounting for statistical uncertainty and allows one to assess and anticipate the status of dynamic processes [7] in real time. In this instance, real-time computed estimates of the covariance matrices of the designated random processes are used to adapt the model to the features of always available random disturbances and measurement noise. The capacity to explicitly consider the statistical properties of measurement noises and state disturbances, the ability to calculate optimal estimates of state variables and their forecasts, the possibility to perform effectively data fusion, the ability to estimate unmeasured components of the state vector, and the capacity to estimate states and some model parameters simultaneously are some of the benefits of available optimal filtering procedures.

In state space format, Kalman filters are used to estimate states based on linear or nonlinear dynamical sys-

DOI 10.15588/1607-3274-2024-2-5

tems. The evolution of the state from time $n-1$ to time n is defined by the process model as follows [8]:

$$x(n) = Ax(n-1) + Bu(n-1) + w(n-1).$$

The process model is paired with the measurement model that describes the relationship between the state and the measurement at the current time step n as:

$$z(n) = Hx(n) + v(n).$$

Given the initial estimate of $x(0)$, the series of measurements, $z(1), z(2), ..., z(n)$, and the details of the system model defined by $A, B, H, Q,$ and R, the task of the Kalman filter is to generate an optimal estimate of $x(n)$ at time n.

In many real-world applications, the true statistics of the noises are either unknown or not Gaussian, despite the fact that the covariance matrices are meant to represent their statistics. As a result, Q and R are typically employed as tuning parameters, which the user can modify to get the intended filter performance.

The covariance matrix of errors of state vector estimates, which is connected to the state estimate, is also used in this technique. It is denoted as $P(n)$.

Algorithm of Kalman filter consists of the next steps.

1. For the state vector and the covariance matrix of estimate errors $P(0)$, set the initial conditions $x(0)$. Assign values to measurement errors R and state disturbances of covariance matrices Q.

2. Determine the filter's matrix optimum coefficient as follows:

$$K(n) = \hat{P}(n-1)H^T [HP(n-1)H^T + R]^{-1}.$$

3. To get the state vector's current estimate, use the new measurements:

$$\tilde{x}(n) = A\tilde{x}(n-1) + K(n)[z(k) - HA\tilde{x}(n-1)].$$

4. For updated estimations, compute the posterior covariance matrix of errors:

$$P(n) = [I - K(n)H]\hat{P}(n).$$

5. Determine the a priori covariance matrix of estimate errors (for the subsequent state vector estimation):

$$\hat{P}(n+1) = AP(n)A^T.$$

then proceed to step 2 (the filter equations subsequent calculation).

The authors of [9] introduced KalmanNet, a hybrid system that combines the traditional model-based extended Kalman filter with deep learning techniques. Their method learns to overcome model mismatches and nonlinearities while enabling real-time state estimation in the same way as model-based Kalman filtering.

The drawbacks of low filtering accuracy and the divergence of conventional nonlinear algorithms in situations when the system noise is unknown can be successfully addressed by the suggested technique in [10]. Additionally, the filter's stability and flexibility are enhanced by the suggested algorithm.

The uncertainty of a financial loss that insurers assess using statistical and mathematical techniques is known as actuarial risk. Actuaries assist insurance firms in correctly setting premiums and reserves by analyzing previous data to estimate future risks. Policyholder protection and financial stability are guaranteed by this delicate balance.

With Generalized Linear Models (GLM), assumptions on the characteristics of the insurance data and how they relate to the anticipated variables can be made explicitly. Moreover, GLM offer statistical diagnostics that support the process of identifying just important variables and validating model hypotheses. This methodology is commonly acknowledged as a conventional approach to insurance pricing across many markets and nations.

As a particular instance among the many models that make up the GLM, there is the linear and nonlinear regression model. Rejecting assumptions for the latter include additive nature of effects, constant variance, and a normal distribution. One possible source for the target variable is an exponential family of distributions [11].

The general form of the exponential family of distributions is as follows:

$$f(y, \theta, \varphi) = \exp\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi\}.$$

Both the variance and the distribution mean may fluctuate. It is expected that explanatory variables have an additive effect on a different scale. For GLM, the following presumptions are made:

1. Stochastic component: every element that makes up y is independent and comes from the same exponential family distribution.

2. Systematic component: the linear predictor η is formed of p covariates, or explanatory variables:

$$\eta = X\beta.$$

3. Link function: a differentiable, monotonic link function establishes the linkage between the random and systematic components.

$$E[y] = \mu = g^{-1}(\eta).$$

Forward stepwise regression produced very good results for the identification of risk variables in [12]. When the technique for selecting risk factors was not used prior

DOI 10.15588/1607-3274-2024-2-5

to inclusion in GLM, it has discovered a number of risk variables for both the frequency and severity of claims, improving the predictive performance of the GLM in comparison to the traditional approach.

It was discovered in [13] that iterative algorithm for generalized linear models with credibility (GLMC) works best when combined with exhaustive variable selection techniques. Its computational efficiency and simplicity enable a rapid estimation of model parameters.

The estimate of GLM parameters is a major issue that has to be given enough consideration in the process of model constructing. The following methods were used successfully to evaluate the parameters: Markov chain Monte Carlo method (MCMC), Adaptive moment estimation (Adam) optimization algorithm, and Weighted least squares iterative-recursive approach (IRWLS).

These algorithms are fully described in the following works [14–16].

4 EXPERIMENTS

Since insurance data is not always accessible to the general public, it was chosen to create target variables and insurance indicators at random using simulation approach. The data structure consists of the next features:

– Age is a numerical variable, which was generated in range from 18 to 64;

– Sex is a categorical string variable;

– Body mass index is numerical variable, which was generated by using normal distribution;

– Number of children is a numerical variable, which was generated in range from 0 to 5;

– Smoker is a categorical string variable;

– Region is a categorical string variable, which was generated from sample: “east”, “south”, “west”, “center”, “north”;

– Charges is a numerical variable.

The target is the final variable, and the distribution laws and matching link functions listed below were applied to it:

– a normal distribution with a logarithmic link function and a known variance σ;

– an exponential distribution using the link function of identity;

– Pareto distribution with a link function of the following type $f(x) = -1 - x$ and a given scale parameter, x_m.

The predicted variable was supplemented with Gaussian noise with varying variance, which is a linear function.

5 RESULTS

After applying Kalman filter original insurance charges were compared with original values.

Results of applying Kalman filter on charges for different distributions is shown on Figures 1, 2, 3.

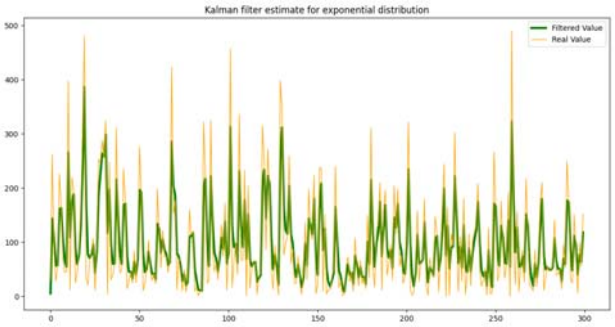

Figure 1 – Results of applying Kalman filter on charges with exponential distribution using the link function of identity

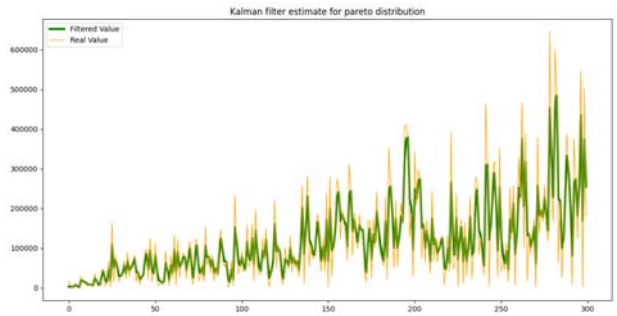

Figure 2 – Results of applying Kalman filter on charges with Pareto distribution with a link function of the negative linear function and a given scale parameter, x_m

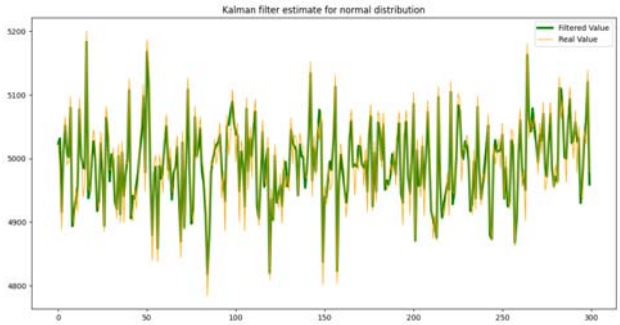

Figure 3 – Results of applying Kalman filter on charges with a normal distribution with a logarithmic link function and a known standard devaition σ

The models’ quality was assessed by using next forecasting metrics: Mean squared error (MSE), Root mean squared error (RMSE) and mean absolute error (MAE).

Tables 1–6 exhibit the findings of the estimate of GLM parameters using the three approaches (IRWSL, ADAM and MCMC) with and without the Kalman filter for three proposed distribution laws with specialized link functions.

Table 1 – Results of GLM construction for charges with Gaussian distribution with known variance and a logarithmic link function without Kalman filter

Metric	MCMC	ADAM	IRWLS
MSE	4408.68	684.63	2457.11
RMSE	64.35	25.10	48.4
MAE	48.76	23.76	48.1

DOI 10.15588/1607-3274-2024-2-5

Table 2 – Results of GLM construction for charges with Gaussian distribution with known variance and a logarithmic link function with Kalman filter

Metric	MCMC	ADAM	IRWLS
MSE	165.33	74.5	218.68
RMSE	12.85	8.63	14.79
MAE	3.957	5.65	11.77

Table 3 – Results of GLM construction for charges with Pareto distribution with known scale parameter and a negative linear link function without Kalman filter

Metric	MCMC r	ADAM	IRWLS
MSE	52724.54	88185.03	638115.4
RMSE	228.51	286.65	797.61
MAE	153.03	205.31	771.25

Table 4 – Results of GLM construction for charges with Pareto distribution with known scale parameter and a negative linear link function with Kalman filter

Metric	MCMC r	ADAM	IRWLS
MSE	7858.747	16808.82	22900.647
RMSE	88.65	129.65	151.33
MAE	63.513	112.12	128.532

Table 5 – Results of GLM construction for charges with an exponential distribution and a an identity link function without Kalman filter

Metric	MCMC	ADAM	IRWLS
MSE	211.38	147.9	288.51
RMSE	14.47	12.18	16.7
MAE	11.05	3.63	13.7

Table 6 – Results of GLM construction for charges with an exponential distribution and a an identity link function with Kalman filter

Metric	MCMC	ADAM	IRWLS
MSE	78.21	54.723	107.478
RMSE	8.84	7.397	10.367
MAE	4.1	2.34	7.147

6 DISCUSSION

Uncertainties for statistical data are factors that have a negative impact on the results of calculations performed at all stages of the data processing process. In this work three approaches were implemented for estimating parameters of GLM with and without the preliminary use of the Kalman filter. It is evident from the GLM building results for the three scenarios mentioned that, for the most part, the Adam technique produced quite decent outcomes. In the case of the Pareto distribution, the MCMC approach also produced positive outcomes. It can be seen that applying Kalman filter for preliminary data processing and fitting model provides for better results of the quality metrics used.

CONCLUSIONS

The problem of minimizing influence of uncertainties in the process of analysis of actuarial risks regarding forecasting charges is solved in this work.

The scientific novelty of obtained results shows that combination of generalized linear models and optimal Kalman filter can be used for building efficient and adequate high quality forecasting models.

The practical significance of current work and its results is that future prospects for further research may include the use of this approach in other fields of insurance related to analysis of actuarial risk.

Prospects for further research are to study the other approaches that can be used to reduce negative influence of possible data and expert estimates uncertainties related to the analysis of actuarial risks. A specialized decision support system should be designed and implemented to solve the problem.

ACKNOWLEDGEMENTS

The author of the presented research results wants to appreciate his scientific advisor, Petro I. Bidyuk – Dr. Tech. Sc., Professor at the Department of Mathematical Methods of System Analysis, Institute for Applied Systems Analysis at the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

REFERENCES

1. Toma S. V. Chiriță M., Șarpe D. Risk and uncertainty, *Procedia Economics and Finance*, Vol. 3, pp. 975–980. DOI: https://doi.org/10.1016/S2212-5671(12)00260-2.
2. Meyer V. R. Measurement uncertainty, *Journal of Chromatography A*, 2007, Vol. 1158, № 1–2, pp. 15–24. DOI: https://doi.org/10.1016/j.chroma.2007.02.082.
3. Laskey K. B., da Costa P. C. G., Tolk A., Jain L. C. (eds) Uncertainty Representation and Reasoning in Complex Systems, *Complex Systems in Knowledge based Environments: Theory, Models and Applications*. New York. Springer, 2009, Ch. 2, pp. 7–40. DOI: https://doi.org/10.1007/978-3-540-88075-2_2.
4. Dyer J. S., McDaniel R. R., Driebe D. J. 15 The Fundamental Uncertainty of Business: Real Options, *Uncertainty and Surprise in Complex Systems: : questions on working with the unexpected*. Berlin, Springer-Verlag, 2005, pp. 153–164. DOI: https://doi.org/10.1007/10948637_15.
5. Makridakis S., Bakas N. Forecasting and uncertainty: A survey, *Risk and Decision Analysis*, 2016, Vol. 6, № 1, pp. 37–64. DOI: http://dx.doi.org/10.3233/RDA-150114.
6. Buchanan M. Forecast: what physics, meteorology, and the natural sciences can teach us about economics. USA, Bloomsbury Publishing, 2013, 272 p.
7. Wu X., Kumar V., Ross Quinlan J. et al. Top 10 algorithms in data mining, *Knowledge and Information Systems*, 2008, Vol. 14, pp. 1–37. DOI: https://doi.org/10.1007/s10115-007-0114-2.
8. Urrea C., Agramonte R. Kalman filter: historical overview and review of its use in robotics 60 years after its creation, *Journal of Sensors*, 2021, Vol. 2021, pp. 1–21. DOI: https://doi.org/10.1155/2021/9674015.
9. Revach G., Shlezinger N., Xiaoyong N. et al. KalmanNet: Neural network aided Kalman filtering for partially known dynamics, *IEEE Transactions on Signal Processing*, 2022, Vol. 70, pp. 1532–1547. DOI: https://doi.org/10.48550/arXiv.2107.10043.
10. Xu D., Wang B., Zhang L. et al. A New Adaptive High-Degree Unscented Kalman Filter with Unknown Process Noise, *Electronics*, 2022, Vol. 11, № 12, pp. 1863–1874. DOI: https://doi.org/10.3390/electronics11121863.
11. Anderson D., Feldblum S., Modlin C. et al. A practitioner’s guide to generalized linear models, *Casualty Actuarial Society Discussion Paper Program*, 2004, Vol. 11, Issue 3, pp. 1–116.
12. Omerašević A., Selimović J. Risk factors selection with data mining methods for insurance premium ratemaking, *Zbornik Radova Ekonomski Fakultet u Rijeka*, 2020, Vol. 38, № 2, pp. 667–696. DOI: https://doi.org/10.18045/zbefri.2020.2.667.
13. Campo B. D. C., Antonio K. Insurance pricing with hierarchically structured data an illustration with a workers’ compensation insurance portfolio, *Scandinavian Actuarial Journal*, 2023, Vol. 2023, Issue 9, pp. 853–884. DOI: https://doi.org/10.1080/03461238.2022.2161413.

DOI 10.15588/1607-3274-2024-2-5

14. McCullagh P., Nelder J. Generalized Linear Models. Second edition. London, Chapman & Hall, 1989, 532 p.
15. Akrout M., Tweed D. On a Conjecture Regarding the Adam Optimizer, 2022. [Electronic resource]. Access mode: https://arxiv.org/pdf/2111.08162.pdf.
16. Roy V. MCMC for GLMMs, 2022. [Electronic resource]. Access mode: https://arxiv.org/pdf/2204.01866.pdf.

Received 23.01.2024.
Accepted 20.04.2024.

УДК 004.852

АНАЛІЗ НЕВИЗНАЧЕНОСТЕЙ ДАНИХ У МОДЕЛЮВАННІ ТА ПРОГНОЗУВАННІ АКТУАРНИХ ПРОЦЕСІВ

Панібратов Р. С. – аспірант Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу аналізу невизначеностей даних у моделюванні та прогнозуванні актуарних процесів. Об'єктом дослідження є задача прогнозування страхових виплат на основі даних про страхових клієнтів з врахуванням можливих ситуацій невизначеності.

Мета роботи – розробка підходу, що дозволяє спрогнозувати майбутні страхові виплати з попередньою мінімізацією можливої невизначеності статистичних даних.

Метод. Запропоновано метод, що дозволяє реалізувати алгоритми оцінювання параметрів узагальнених лінійних моделей з попереднім використанням оптимального фільтру Калмана. Результати продемонстрували більш якісні результати прогнозу та більш адекватні структури моделі. Даний підхід був успішно застосований на процедурі штучно згенерованих страхових даних. Для генерування страхового набору даних клієнтів були використані наступні показники: вік; стать; індекс маси тіла (використовуючи нормальний закон розподілу); кількість дітей (від 0 до 5); статус курця; регіон (північ, схід, південь, захід, центр); виплати. Для створення останньої величини використовувався нормальний розподіл з відомою дисперсією і логарифмічною функцією зв'язку, експоненційний розподіл з одиничною функцією зв'язку та розподіл Парето з відомим параметром масштабування і від'ємною лінійною функцією зв'язку.

Результати. Запропонований підхід реалізований програмно у вигляді системи обробки інформації для розв'язування задачі прогнозування страхових виплат за страховими даними та з урахуванням зашумленості даних.

Висновки. Запропонований підхід реалізований програмно для побудування більш адекватних моделей та розв'язування задачі точного прогнозування страхових виплат за страховими даними та врахуванням зашумленості даних. Перспективи подальших досліджень можуть включати використання даного підходу в інших областях застосування, що пов'язані з акутарним ризиком. Необхідно розробити спеціалізовану інтелектуальну систему підтримки прийняття рішень для розв'язування задач з використанням страхових даних реального світу в режимі онлайн, а також сучасних інформаційних технологій та інтелектуального аналізу даних.

КЛЮЧОВІ СЛОВА: актуарний ризик, узагальнені лінійні моделі, оптимальний фільтр Калмана, експоненційна множина розподілів, моделювання, ітеративно-рекурентно зважуваний метод найменших квадратів, метод Adam, метод Монте-Карло для марківських ланцюгів.

ЛІТЕРАТУРА

1. Toma S. V. Risk and uncertainty. / S. V. Toma, M. Chiriță, D. Șarpe // Procedia Economics and Finance. – Vol. 3 – P. 975–980. DOI: https://doi.org/10.1016/S2212-5671(12)00260-2.
2. Meyer V. R. Measurement uncertainty / V. R. Meyer // Journal of Chromatography A. – 2007. – Vol. 1158, № 1–2. – P. 15–24. DOI: https://doi.org/10.1016/j.chroma.2007.02.082.
3. Laskey K. B. Uncertainty Representation and Reasoning in Complex Systems / Laskey K. B., da Costa P. C. G., Tolk A., Jain L. C. (eds) // Complex Systems in Knowledge based Environments: Theory, Models and Applications. – New York : Springer, 2009. – Ch. 2 – P. 7–40. DOI: https://doi.org/10.1007/978-3-540-88075-2_2.
4. Dyer J. S. 15 The Fundamental Uncertainty of Business: Real Options / J. S. Dyer, R. R. McDaniel, D. J. Driebe // Uncertainty and Surprise in Complex Systems: : questions on working with the unexpected. – Berlin : Springer-Verlag, 2005. – P. 153–164. DOI: https://doi.org/10.1007/10948637_15.
5. Makridakis S. Forecasting and uncertainty: A survey / S. Markidakis, N. Bakas // Risk and Decision Analysis. – 2016. – Vol. 6, № 1. – P. 37–64. DOI: http://dx.doi.org/10.3233/RDA-150114.
6. Buchanan M. Forecast: what physics, meteorology, and the natural sciences can teach us about economics. / M. Buchanan – USA : Bloomsbury Publishing, 2013 – 272 p.
7. Top 10 algorithms in data mining / [X. Wu, V. Kumar, J. Ross Quinlan et al.] // Knowledge and Information Systems. – 2008. – Vol. 14. – P. 1–37. DOI: https://doi.org/10.1007/s10115-007-0114-2.
8. Urrea C. Kalman filter: historical overview and review of its use in robotics 60 years after its creation / C. Urrea, R. Agramonte // Journal of Sensors. – 2021. – Vol. 2021. – P. 1–21. DOI: https://doi.org/10.1155/2021/9674015.
9. KalmanNet: Neural network aided Kalman filtering for partially known dynamics / [G. Revach, N. Shlezinger, N. Xiaoyong et al.] // IEEE Transactions on Signal Processing. – 2022. – Vol. 70. – P. 1532–1547. DOI: https://doi.org/10.48550/arXiv.2107.10043.
10. A New Adaptive High-Degree Unscented Kalman Filter with Unknown Process Noise / [D. Xu, B. Wang, L. Zhang et al.] // Electronics. – 2022. – Vol. 11, № 12 – P. 1863–1874. DOI: https://doi.org/10.3390/electronics11121863.
11. A practitioner's guide to generalized linear models / [D. Anderson, S. Feldblum, C. Modlin et al.] // Casualty Actuarial Society Discussion Paper Program. – 2004. – Vol. 11, Issue 3. – P. 1–116.
12. Omerašević A. Risk factors selection with data mining methods for insurance premium ratemaking / A. Omerašević, J. Selimović // Zbornik Radova Ekonomski Fakultet u Rijeka. – 2020. – Vol. 38, № 2. – P. 667–696. DOI: https://doi.org/10.18045/zbefri.2020.2.667.
13. Campo B. D. C. Insurance pricing with hierarchically structured data an illustration with a workers' compensation insurance portfolio / B. D. C. Campo, K. Antonio // Scandinavian Actuarial Journal. – 2023. – Vol. 2023, Issue 9 – P. 853–884. DOI: https://doi.org/10.1080/03461238.2022.2161413.
14. McCullagh P. Generalized Linear Models. Second edition / P. McCullagh, J. Nelder. – London : Chapman & Hall, 1989. – 532 p.
15. Akrout M. On a Conjecture Regarding the Adam Optimizer, 2022. [Electronic resource] / M. Akrout, D. Tweed. – Access mode: https://arxiv.org/pdf/2111.08162.pdf.
16. Roy V. MCMC for GLMMs, 2022. [Electronic resource] / V. Roy. – Access mode: https://arxiv.org/pdf/2204.01866.pdf.

DOI 10.15588/1607-3274-2024-2-5