VALIDATION FRAMEWORK FOR ADVANCED FINANCIAL FORECASTING MODELS IN INSURANCE

Nikolay Pavlov, Zlatomila Mincheva, Maria Dobreva

Abstract. *This research establishes a validation framework for a novel financial forecasting model in insurance, with a primary focus on the critical data preparation stage. We investigate two distinct feature engineering paradigms for handling granular, per-policy data: 1) Aggregation into structured time-series suitable for recurrent or convolutional models, and 2) Direct use of policy-level data streams to leverage sequence-aware architectures like Transformers. The core contribution is the development of this data-processing pipeline and the subsequent rigorous validation of our model. The study provides a critical assessment of the model’s performance, stability, and its dependency on the chosen data representation strategy.*

Key words: Predictive Modeling, Insurance Finance, Model Validation, Time-Series Forecasting, Deep Learning, LSTM, Transformers, Data Engineering.

Introduction

Premium forecasting is a central task in insurance finance, supporting pricing, budgeting, reserving and capital planning. In many applications, forecasts are still derived from aggregated development triangles at portfolio level, where premiums or claims are summarised by origin year and development month and projected using development factors as in classical chain-ladder style approaches [1]. These methods are simple and transparent, but they operate on portfolio-level aggregates and ignore the detailed dynamics of individual policies.

Modern information systems record granular per-policy transactions at monthly resolution or finer. This enables more expressive forecasting models, including deep learning approaches built on origin–development (OD) structures such as DeepTriangle [3]. At the same time, such models are extremely sensitive to how raw operational data are cleaned, structured and transformed.

In this paper we reverse the usual emphasis. Instead of focusing on model architecture, we concentrate on the data-engineering layer that transforms raw certificate-level transactions into clean policy chains, OD tables and origin triangles. The proposed pipeline is model-agnostic: it can feed both traditional actuarial techniques and modern machine-learning models, and serves as the backbone of a validation framework for any forecasting approach applied on the portfolio.

In parallel, a rich body of work has emerged on machine learning and deep learning models for pricing and reserving that operate on triangle or micro-level development data [2, 4, 6]. Our contribution is complementary: we focus on the data-engineering layer that produces consistent OD structures from fragmented policy transactions.

Challenges in Working with Real Insurance Data

These difficulties motivate a carefully designed, policy-level OD pipeline that explicitly reconstructs monthly development before any modelling or validation is attempted.

Real insurance data deviate substantially from the idealised conditions assumed in textbook examples of development triangles [1, 5]. A first difficulty is the fragmentation of policy history. A single policy may generate multiple certificates due to endorsements, mid-term adjustments, corrections or cancellations, often coming from different operational systems. Reconstructing a coherent monthly premium path requires reliable linking across these sources.

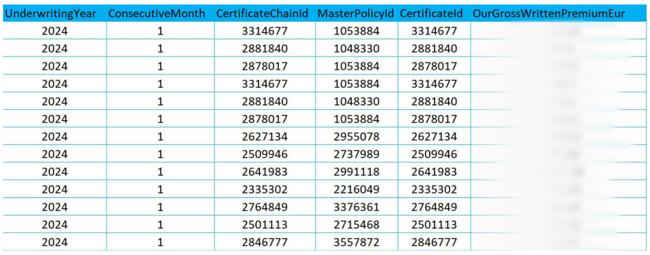
A second difficulty is the presence of missing and inconsistent months. Policies that should be active for a full year frequently exhibit gaps or overlapping certificates for the same month, and temporal fields such as underwriting year, inception month and consecutive month index may be misaligned. Business-driven irregularities add further noise: short-term contracts, early cancellations, seasonal products and one-off corrections create highly irregular development patterns which, when aggregated directly into triangles, produce unstable cells and hard-to-interpret development factors. Finally, the evolution of systems and product definitions introduces structural breaks that are not always documented in a way useful for analytics.

Research Objectives

The research reported in this paper has three tightly related objectives. The first is to design a robust origin–development data-processing pipeline that starts from raw certificate-level transactions and produces a consistent policy–month view, policy chains aligned by origin and development, and OD tables and origin triangles suitable for both actuarial and machine-learning models. A second objective is to demonstrate how this pipeline underpins a validation framework by enabling time-consistent train–test splits, traceability from triangle cells back to individual policies and systematic diagnostics of data quality issues. The third objective is to integrate the OD layer into the existing reporting and business-intelligence infrastructure of the insurer, so that model outputs can be embedded in operational dashboards and automatic reports without duplicating data flows.

Data and Origin–Development Pipeline

The portfolio used in this study is a non-life insurance book observed monthly. For each certificate-level transaction we observe policy and certificate identifiers, underwriting year, consecutive month and gross written premium in euro. After basic validation and de-duplication, the cleaned input appears as in Figure 1, which shows one row per certificate and month together with the aggregated premium amount.



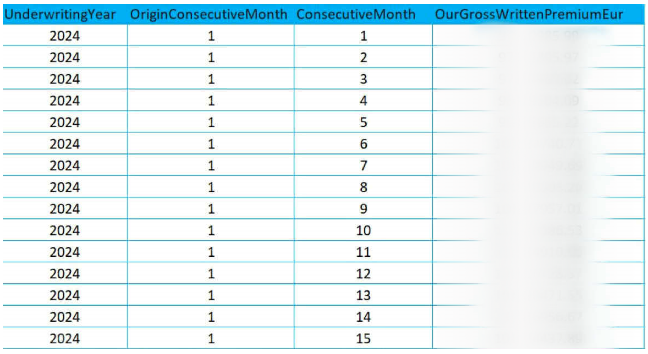
UnderwritingYear	ConsecutiveMonth	CertificateChainId	MasterPolicyId	CertificateId	OurGrossWrittenPremiumEur
2024	1	3314677	1053884	3314677	[redacted]
2024	1	2881840	1048330	2881840	[redacted]
2024	1	2878017	1053884	2878017	[redacted]
2024	1	3314677	1053884	3314677	[redacted]
2024	1	2881840	1048330	2881840	[redacted]
2024	1	2878017	1053884	2878017	[redacted]
2024	1	2627134	2955078	2627134	[redacted]
2024	1	2509946	2737989	2509946	[redacted]
2024	1	2641983	2991118	2641983	[redacted]
2024	1	2335302	2216049	2335302	[redacted]
2024	1	2764849	3376361	2764849	[redacted]
2024	1	2501113	2715468	2501113	[redacted]
2024	1	2846777	3557872	2846777	[redacted]

Figure 1. Example of cleaned policy–month input data

The OD pipeline begins with cleaning and standardisation. Invalid, duplicated or technically inconsistent records are removed or corrected according to business rules. Currencies, date formats and identifiers are harmonised, overlapping certificates for the same policy and calendar month are consolidated into a single net premium, and records that contradict product definitions or accounting rules are excluded. The output of this stage is a consistent policy–month view, where each row describes a unique combination of policy, calendar month and gross written premium.

In the next stage policy chains are constructed. For each policy we identify its origin month, defined as the first month with non-zero premium, and order all subsequent monthly premiums by a development index. Portfolio-specific decisions are made regarding the maximum development horizon, the treatment of missing months as either explicit zeros or flagged missing values, and the handling of lapses, reinstatements and mid-term cancellations. Each policy is thus represented by a sequence of monthly premiums starting at origin and continuing over the chosen horizon.

Once policy chains are available, each policy–month record is mapped to origin–development coordinates consisting of origin year, origin month and development month. Figure 2 illustrates a fragment of the resulting table, which now contains one row per origin–development combination together with the corresponding premium.



UnderwritingYear	OriginConsecutiveMonth	ConsecutiveMonth	OurGrossWrittenPremiumEur
2024	1	1	[redacted]
2024	1	2	[redacted]
2024	1	3	[redacted]
2024	1	4	[redacted]
2024	1	5	[redacted]
2024	1	6	[redacted]
2024	1	7	[redacted]
2024	1	8	[redacted]
2024	1	9	[redacted]
2024	1	10	[redacted]
2024	1	11	[redacted]
2024	1	12	[redacted]
2024	1	13	[redacted]
2024	1	14	[redacted]
2024	1	15	[redacted]

Figure 2. Origin–development level table constructed from policy chains

From this OD table, premiums are aggregated by origin year (or origin year–month) and development month to obtain origin triangles that are structurally comparable to classical actuarial triangles [1, 5]. Because each triangle cell is computed as a sum over explicitly known policy–month contributions, any irregularity in the triangle can be traced back to the underlying set of policies and their development paths.

Role in Model Validation and Reporting

The OD pipeline forms a common data layer for different forecasting models and validation procedures. Origin year and origin month provide a natural chronological index, so models can be trained on older origin years and tested on later ones, mimicking real deployment and avoiding information leakage. Because all models receive inputs derived from the same OD table, differences in performance can be attributed to modelling choices rather than to inconsistent preprocessing. Classical chain-ladder estimators, generalised linear models on triangle cells and deep learning models inspired by DeepTriangle [3] can therefore be evaluated on a common footing.

The same OD layer supports diagnostics of both data and model stability. Repeated refitting of models on rolling origin windows highlights how sensitive results are to changes in business mix or data quality, while anomalous triangle cells or unstable development factors can be related back to specific cohorts of policies via the policy chains. From an implementation point of view, the OD table is integrated into the insurer's Framework for Distributed Business Applications, which already supports automated report generation and scheduling as well as embedded BI dashboards via Power BI [7, 8]. This allows model results based on OD data to appear alongside operational key performance indicators in standard reporting channels.

Advantages and Limitations

The OD-centric approach offers several advantages for practitioners. It provides transparency and traceability, since there is an explicit chain from raw transactions to policy chains, OD table and triangles, and analysts can always reconstruct which policies and months contribute to a given cell. It forces explicit handling of data defects, because missing months, short histories and structural breaks are visible at OD level, where they can be quantified and, if necessary, filtered or down-weighted before modelling. It also creates a model-agnostic foundation: the same OD table can serve as input to traditional actuarial techniques, gradient-boosting models or deep neural networks, enabling a consistent validation framework irrespective of model class.

At the same time, the approach has limitations. It requires detailed per-policy transactional data and cannot be applied directly in organisations that store only aggregated figures. Policies with extremely short or idiosyncratic histories remain difficult to encode and may need special treatment rules. Building and maintaining the OD pipeline demands close collaboration between actuaries, data engineers and business experts to ensure that technical transformations respect product logic and regulatory constraints.

Although we do not develop a specific forecasting model here, the OD table constructed by our pipeline can be used directly as input to modern reserving and pricing approaches of the type discussed in [2, 4, 6].

Conclusion

This paper presented a data-centric origin–development pipeline for per-policy premium data and its role in a validation framework for advanced insurance forecasting models. Starting from fragmented certificate-level transactions, the pipeline constructs a clean policy–month view, policy chains, an OD table and origin triangles that remain fully traceable back to individual policies. Rather than prescribing a specific model, the framework emphasises that data preparation and OD structuring are prerequisites for any serious model validation exercise. By providing a stable and transparent data layer, the pipeline enables fair comparison of heterogeneous models, systematic diagnostics of data quality issues, and straightforward integration with existing BI and reporting infrastructure.

Future work will extend the OD pipeline to other product lines and to claims data, refine the treatment of incomplete and short policy chains, and combine OD-based validation with synthetic data generation for stress-testing forecasting models under alternative portfolio scenarios [9].

Acknowledgments

This study is supported by the project FP25-FMI-010 "Innovative interdisciplinary research in informatics, mathematics and educational pedagogy" at the Paisii Hilendarski University of Plovdiv.

References

[1] T. Mack, Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates, *ASTIN Bulletin*, 1993, 23 (2), pp. 213–225

[2] C. Blier-Wong, H. Cossette, L. Lamontagne, É. Marceau, Machine Learning in P&C Insurance: A Review for Pricing and Reserving, *Risks*, 2021, 9 (1), Article 4

[3] K. Kuo, DeepTriangle: A Deep Learning Approach to Loss Reserving, *Risks*, 2019, 7 (3), Article 97

[4] I. Chaoubi, C. Besse, H. Cossette, M. Côté, Micro-Level Reserving for General Insurance Claims Using a Long Short-Term Memory Network, *Applied Stochastic Models in Business and Industry*, 2023, 39, pp. 1–26

[5] E. Frees, *Loss Data Analytics*, Open Actuarial Textbooks, Madison, 2018

[6] M. Wüthrich, Machine Learning in Individual Claims Reserving, *Scandinavian Actuarial Journa*l, 2018, 2018 (6), pp. 465–480

[7] N. Pavlov, M. Dobreva, A. Rahnev, G. Spasov, Automatic Report Generation in FDBA, *Scientific Conference "Innovative ICT in Business and Education: Future Trends, Applications and Implementation"*, Pamporovo, Bulgaria, 24–25 November 2016, pp. 21–28

[8] M. Dobreva, N. Pavlov, A. Rahnev, Integrate Power BI with WPF Desktop Applications, *Scientific Conference "Innovative ICT in Research and Education: Mathematics, Informatics and Information Technologies"*, Pamporovo, Bulgaria, 29–30 November 2018, pp. 6–72

[9] S. Monov, Z. Mincheva, N. Pavlov, Synthetic Time Series Data in Restaurants Supply Chain Planning, *International Scientific Conference IMEA'2024*, Pamporovo, Bulgaria, 13–15 November 2024, pp. 143–148

Nikolay Pavlov[1], Zlatomila Mincheva[1], Maria Dobreva[1]
[1] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria
Corresponding author: nikolayp@uni-plovdiv.bg