



Occurrence and reporting dynamics of weather-related home insurance claims

Due to climate change, insurers are increasingly exposed to weather-related claims in their home insurance portfolios. Modelling the occurrence dynamics of these claims traditionally starts from an identified severe weather event, such as a wind- or hailstorm, after which the claims caused by that event are analysed. In practice, however, claims do not only arise on the handful of days with severe weather. Moderate conditions generate fewer claims per day, but there are many such days, and together they account for a substantial share of the total. Additionally, the claims handling process runs continuously. An insurer therefore needs insight into the occurrence and reporting of claims on every calendar day, not only on the days that make the news. Such insight serves several actuarial purposes: estimating the number of outstanding claims for reserving, planning the capacity of claim handling teams, and, given reliable forecasts, warning policyholders ahead of severe weather so that some

claims are prevented. Two complications arise when all calendar days are modelled jointly. First, claims that occurred recently have only been partially reported, so the most recent observations are incomplete. Second, the rare days with severe weather are pooled with the many moderate days, while the dynamics on those days differ substantially. Using machine learning techniques, we model the occurrence and reporting of weather-related home insurance claims per calendar day, accounting for the claims not yet reported and allowing the dynamics to differ across weather conditions.

DATA

We have access to a unique Dutch home insurance portfolio and combine this with publicly available weather data. The portfolio consists of over one million insured objects, with weather-related claim information from January 2017 to December 2022. For each object, we know its characteristics, such as roof type, object type and building year, and the PC4 postal code area in which it is located. For each calendar day within the observed period, we construct day-specific covariates such as the day of the week and indicators for Dutch national holidays. We collect weather data from three public sources: the ERA5 reanalysis and the E-OBS observational dataset, both available via the Copernicus Climate Data Store, and radar- and station-based datasets from the KNMI. These sources cover wind speed, precipitation, hail probability, temperature and surface pressure, and differ considerably in coverage, grid granularity and temporal resolution. The weather data are defined on grids that do not align with the postal

code areas in which the insured objects are located, while the object characteristics are available for each individual object. An extensive feature engineering step is therefore required to bring all sources to a common spatial and temporal resolution. Part of this step exploits the hourly weather measurements to construct daily covariates that capture the duration and the intra-day timing of severe weather, for instance by calculating the number of hours during which wind speed exceeds preselected thresholds.

Figure 1 illustrates why this data combination is informative. It shows the precipitation recorded on 4 June 2019, a day with a rain- and hailstorm, alongside the postal code areas in which claims occurred. The band of severe weather running from the south of the Netherlands to the north-east is clearly visible in both panels, and would be averaged away entirely if the whole country were treated as a single area.

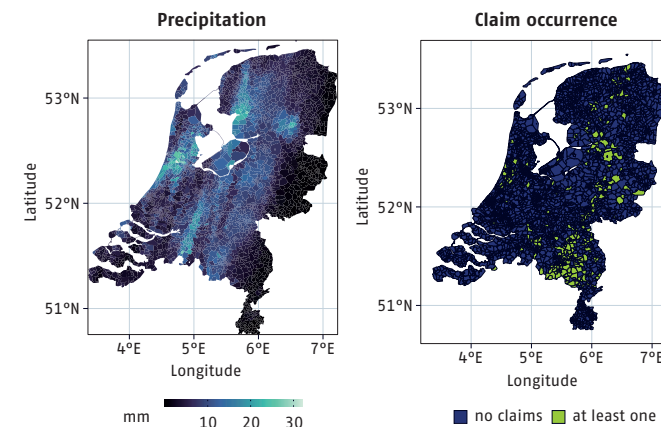


Figure 1: Total daily precipitation (left) and the PC4 postal code areas in which at least one claim occurred (right), for the rain- and hailstorm of 4 June 2019.

MODELLING APPROACH

We model, for each combination of a calendar day and a postal code area, the number of claims that occur and the delay with which these claims are reported. Both components are estimated with extreme gradient boosting, a tree-based machine learning technique which handles the large number of weather covariates and the strong correlations between them. It automatically captures non-linear effects and interactions without requiring us to specify them in advance. A complication is that claims which occurred recently have not all been reported yet, so the observed claim counts for the most recent days understate what actually happened. We address this with an expectation-maximisation algorithm, which alternates between estimating the number of claims that are still outstanding and refitting the occurrence and reporting models on the estimated and observed data. On top of this, we let the dynamics differ across weather conditions. Rather than a priori defining severe weather events, we learn a regime structure from the data itself by applying a clustering algorithm to the weather data. We then fit a separate occurrence model within each weather regime. We use feature importance to identify which covariates drive the occurrence and the reporting of claims, and compare predictive performance for the number of outstanding claims.

RESULTS AND INSIGHTS

We compare a single-regime and a two-regime specification at two spatial resolutions: the Netherlands as a whole, and its 90 PC2 postal code areas. At the regional level, the two-regime model outperforms the single-regime model on every criterion. At the national level, it improves the reporting fit but worsens the occurrence fit. We find that a multi-regime structure adds most value when the data are modelled at a granular spatial resolution. When the entire country is treated as a single area, each day corresponds to one averaged weather profile, so a storm hitting only part of the Netherlands cannot be captured. At the regional level, the same day can fall in the severe weather regime in one part of the country and in the clear weather regime in another.

Figure 2 shows the daily difference between the predicted and the observed number of claims, for the national single-regime model and the regional two-regime model. The gap between the two is driven by severe windstorms: the national single-regime model substantially underestimates the claim counts during the windstorms Friederike (18 January 2018) and Eunice (18 February 2022), compared to the regional two-regime model.

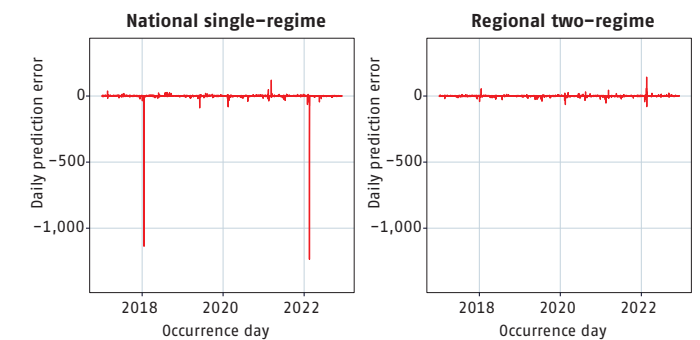


Figure 2: Daily prediction error, i.e. the difference between the predicted and observed number of claims, for each occurrence day.

Another finding concerns the covariates constructed from hourly weather measurements. Features capturing the duration of severe weather, such as the number of hours during which wind speed exceeds a threshold, and its intra-day timing, such as whether wind intensity peaks in the evening, are informative for the occurrence of claims as well as for their reporting. Damage late in the evening is largely reported the following day rather than on the day of occurrence itself. What matters for the reporting model is therefore not only how severe the weather for a given day was, but when within that day the severe weather occurred. A model with only daily maxima and averages cannot distinguish two such days, even though their reporting patterns differ considerably.

CONCLUSION

This study develops a framework that models the occurrence and reporting of weather-related home insurance claims on every calendar day, combining portfolio information with publicly available weather data. In a case study on a Dutch portfolio, we find that a multi-regime structure is most suitable when sufficiently granular spatial data is available, and that covariates capturing the duration and intra-day timing of severe weather improve both the occurrence and the reporting model. For an actuary, our model can improve the estimation of outstanding claims for reserving, while supporting better capacity planning for claim handling teams. Combined with reliable weather forecasts, it also opens the door to warning policyholders ahead of severe weather, shifting the insurer's role from paying for damage to helping avoid it. Insurers and society have the same interest here: a claim prevented is better than a claim paid. Sharing early warnings with public authorities is a natural next step, and one where research of this kind can help insurers contribute well beyond their own portfolios. ■

From left to right:

P. Wilsens MSc is a PhD student at LRISK, the insurance research group within the Faculty of Economics and Business (FEB), KU Leuven.
Prof. dr. K. Antonio is full professor in actuarial science at KU Leuven (Belgium) and part time professor in actuarial data science with the University of Amsterdam (the Netherlands).
Prof. dr. G. Claeskens is full professor in statistics at KU Leuven (Belgium).
Dr. K.B. Gubbels AAG is a senior actuary at Achmea focusing on market and catastrophe risks.

