TIME SERIES FORECASTING OF OCCUPATIONAL INJURY RISK

GIUSEPPE COCOMAZZI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

547014

COMMITTEE

dr. Stijn Rotman
dr. Travis Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

December 22, 2025

WORD COUNT

4782

ACKNOWLEDGMENTS

TIME SERIES FORECASTING OF OCCUPATIONAL INJURY RISK

GIUSEPPE COCOMAZZI

Abstract

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The SIR OSHA dataset is publicly available at Severe Injury Dashboard | Occupational Safety and Health Administration. The data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.

All the figures belong to the author. In terms of writing, a generative language model, OpenAI's ChatGPT 5.1, was used to improve the author's original content, for paraphrasing, spell checking and grammar, and help debug LaTex syntax. The same model assisted in writing and documenting the code, included the function to fetch data from the U.S. Bureau of Statistics. No other typesetting tools or services were used.

2 PROBLEM STATEMENT & RESEARCH GOAL

2.1 *Context*

Occupational safety and health (OSH) organizations across the world help foster the well-being of workers by preventing workplace accidents, promoting safe practices, and enforcing regulatory standards. Employers are legally responsible for ensuring a healthy working environment, as employees may be exposed to a wide range of physical, biological, and psychosocial hazards.

According to the International Labour Organization (ILO), an estimated 364 million workers globally suffer non-fatal occupational accidents each year, with substantial geographical disparities in incidence rates. Injured

workers often face long-term health complications, reduced employability, and financial hardship.

Beyond its societal implications, workplace safety has a considerable economic dimension. It is widely recognized that improving workplace safety can reduce economic losses at both firm and national levels. ILO estimates suggest that almost an annual 4% loss in global Gross Domestic Product is accountable to occupational accidents and diseases (Organization, 2020).

Traditional prevention measures to reduce occupational injury risk include compliance with safety standards and regulatory frameworks; inspections and audits; environmental controls and maintenance schedules; safety education programs, and post-incident analysis. Despite the efforts, manual reporting increasingly struggles to keep up with both rapidly evolving working conditions and the increasing volume of historical safety data. Post-hoc interventions are no longer sufficient for effective prevention.

For these reasons, real-time responsiveness has become crucial to further mitigate workplace injury risks. The integration of the Internet of Things technologies into the workplace allows for a continuous streaming of data coming from sensors, wearables, and smart equipment, supporting constant monitoring of safety conditions and triggering timely preventive actions (Michaels & Wagner, 2025).

From a scientific point of view, this complex data flow may pose a challenge to traditional time series methods such as Error, Trend, Seasonal models (ETS), autoregressive integrated moving average (ARIMA) and its seasonal extension (SARIMA), which have long served as standard baselines for forecasting tasks in the public-health domain. These models provide statistically grounded forecasts that often perform well in settings with clear seasonal structure and relatively smooth temporal dynamics. On the other hand, Machine learning (ML) methods can leverage rich and dynamical data sources to capture complex, non-linear patterns, thereby marking a shift from a reactive to a proactive safety management strategy (Kontopoulou, Panagopoulos, Kakkos, & Matsopoulos, 2023; Masini, Medeiros, & Mendes, 2023). Because the temporal factor will be the main leading axis in our analysis, we will first explore the data to extract components such as trend and seasonality to guide our model selection. Additionally, we will explore the geographical distribution as well to test whether encoding states at different levels can enhance the interpretability of the models and improve the forecasts both at the local and the global level (respectively, state and national level).

While most studies have focused on classification tasks, the area of time series forecasting for occupational injury prevention remains still

underexplored, despite its potential to anticipate future risk patterns and support more targeted interventions.

2.2 *Research strategy*

As outlined in subsection 2.1, knowing *when* an intervention may be required is crucial for increasing the likelihood of alleviating workers' exposure to severe injuries. If meaningful patterns in historical data can be reliably captured by forecasting algorithms, stakeholders can make more informed decisions about how to prioritize preventive resources, and about which types of data need continuous monitoring. As different temporal granularities can impact the accuracy of the generated forecasts (Rostami-Tabar, Goltsos, & Wang, 2023), the first question we aim to explore is to what extent different levels of temporal resolution can help smooth out an intermitting time series at the original lead time. Moreover, this question gains a further dimension when the data is organised according to a grouping factor, such as having time series across multiple states (Qu, Timmermann, & Zhu, 2024). The model choice will take into account the explainability of the models. **RQ1**: *How does the forecasting performance of a set of classical statistical models (Holt-Winter method, ARIMA, and SARIMA) and ML models (divided into tho main families: linear and tree-based ensemble models) at two temporal aggregations (monthly vs. weekly) in one-step-ahead forecasting of severe injuries risk across a twelve-month horizon, as evaluated by RMSE and MAE?*

This question addresses to what extent the temporal resolution affects the model performance when data are aggregated at a national level, yeilding a single time series. Weekly aggregation increases the number of observations but it may introduce noise and variability, while monthly aggregation may produce coarser predictions. Understanding this trade-off is essential for designing reliable predictive pipelines. This question serves as a general baseline to uncover trend and seasonality patterns and to test whether

Given the absence of a comparable supervised regression setting in prior work on occupational injury forecasting, we will train different families of regressors based on the exploratory initial step and on the main findings from related datasets. The choice for linear models with penalty terms (Ridge, Lasso, ElasticNet) allows us to assess how regularization can affect feature selection in a strongly seasonal data structure across a balanced panel (Han, Zhou, Sun, & Li, 2025). Among tree-based ensemble models, CatBoost is deemed particularly suited for our panel data, as it can

handle categorical features (Hall & Rasheed, 2025). XGBoost and LightGBM appears to consistently be the most accurate models in recent forecasting benchmarks and machine-learning competitions (Makridakis, Spiliotis, & Assimakopoulos, 2022). The sample size and a low-dimensional feature matrix allowed us to compare a wide variety of models with negligible computational cost.

RQ2: *How does the predictive accuracy of the best-performing model change when estimated as a Global Forecasting Model (across all states jointly) rather than as a Local Forecasting Model (state-specific), for both monthly and weekly temporal resolutions?*

With this question we aim to evaluate whether pooling information across states enhances the predictive performance relative to modeling each state independently. A global model may exploit shared temporal knowledge and increase the available sample size, while a local model may better capture state-specific patterns.

SQ2.1: How does a SARIMAX benchmark compare to the best-performing Local Forecasting Model on the same monthly forecast horizon, as evaluated by RMSE and MAE on the target hospitalization?

Traditional statistical models such as SARIMAX can not estimate parameters globally across a panel. Therefore, SARIMAX serves as an appropriate baseline for comparison against local machine-learning models.

RQ3: *Which combinations of lookback windows and forecast horizons minimize out-of-sample forecast errors (RMSE and MAE)?*

Since design choices can affect the forecasting accuracy, we investigate a range of configurations to identify which one yields the most accurate predictions.

2.3 *Societal Relevance*

OSH agencies must strategically allocate their resources to maximize injury reduction. Severe injuries, in particular, can pose an even higher burden for both employees and the society at large. Forecasting the relative risk of these life-altering events in the area where they will most likely occur can provide actionable insights to guide OSH agencies on how to better target their interventions.

Recent evidence underscores the potential of such data-driven targeting strategies. As Johnson, Levine, and Toffel (2023) suggest, a machine learn-

ing approach to prioritize inspections could have averted twice as many injuries compared to the current OSHA prevention programs. Building on this premise, the present study focuses on forecasting relative risk trajectories across U.S. states to support more informed and proactive resource deployment.

2.4 *Scientific Relevance*

A debated topic in the forecasting literature concerns to what extent global models – trained on pooled cross-sectional time-series – can outperform local models, which are trained independently for each unit (e.g., state, firm, or region). We systematically test whether a global pooling strategy improves forecasting accuracy compared to a local strategy when applied to a real-world panel dataset. Multiple ML regressors will be trained and compared under both frameworks, providing empirical evidence on the benefits and limitations, if any, of pooling information across states. Additionally, by framing the problem as a supervised regression task, we incorporate lagged features, rolling statistics, and exogenous variables, in line with the best practice in time series forecasting with ML regressors (Bojer, 2022; Makridakis et al., 2022).

To the best of our knowledge, this is the first study to comparatively assess global versus local forecasting strategies for occupational injury risk prediction. More specifically, although time series forecasting is well established in adjacent health-related domains such as epidemiology and biostatistics, the regression-based framing of our study extend methodological practices common in these scientific fields to occupational safety research (Reich et al., 2019).

3 RELATED WORK

3.1 *Occupation Injury Risk Prediction*

The field of occupation injury risk prediction using ML models has been largely dominated by classification tasks. In their comprehensive survey, Vivian, Bauder, and Khoshgoftaar (2025) identified three main research directions: (i) identifying workplace injury risk factors, (ii) predicting return-to-work rates, and (iii) analyzing sociodemographic features across different worker populations. Most studies rely on business-level datasets or focus on specific injury types such as traumatic brain injuries or foot and ankle injuries.

The ML models most frequently used in this field are selected for their ability to capture patterns in heterogeneous sets of categorical vari-

ables. While no single model clearly outperforms the others, Random Forest, Support Vector Machines, Decision Trees, and Linear Regression have achieved the best results across evaluation metrics such as accuracy, sensitivity, and F-1 score, underscoring the importance of tailoring model selection to data characteristics. However, regression-based approaches that model the temporal evolution of occupational risk remain virtually absent from the literature. Only one study integrated time-series forecasting techniques within a broader classification model to forecast low-severity injuries (Rahman, Hossain, & Sikder, 2024).

The work most closely aligned with our research is Cerqua, Giannantoni, Letta, and Pinto (2024). The authors develop a supervised regression framing for the time-series forecasting of workplace fatalities in Italy, focusing on spatial heterogeneity and ex-ante allocation of regulatory interventions. The methodology explicitly addresses how to apply machine learning pipelines to panel data, implementing a rolling forecasting origin approach for hyperparameter tuning, and evaluating model performance on a one-year hold-out test set. They selected a mix of models: linear, tree-based, and the Long Short-Term Memory neural network, ultimately identifying Partial Least Squares (PLS) as the best performing model across different sets of predictors.

3.2 *Studies on OSHA's SIR dataset*

A notable study examining the temporal aspects of occupational injury data is Gomes, Parasram, Collins, and Socias-Morales (2023). The authors analysed the Severe Injury Report (SIR) dataset provided by the Occupational Safety and Health Administration (OSHA), using time series decomposition models to uncover seasonal effects and general trends. Moreover, they evaluated industry and injury-type breakdowns, revealing that manufacturing and construction are the most affected industries. Importantly, they identified systematic temporal patterns in the occupational injuries, such as the predominance of injuries on weekdays compared to weekends, and a higher injury incidence in the summer months. While their analysis provided valuable descriptive insights into the temporal features of the SIR data, it did not extend to predictive or regression-based modeling.

On the classification side, Khairuddin et al. (2022) trained five machine learning models on the same SIR dataset to predict hospitalization and amputation outcomes, finding that Random Forest achieved the highest performance on metrics such as accuracy and F1-score. Building on this work, Hasan Khalleefah Hassan and Khalifa (2025) evaluated a different set of machine learning models, and reported that the AdaBoost Classifier showed the best performance on accuracy and F1-score. As both studies

aimed to categorize past events, their results and model selection can not serve as a benchmark for our work, but only as a point of reference. In fact, a regression-based forecasting approach applied to SIR data is still absent from the current literature.

3.3 *Theoretical Framework*

When observations are collected over time from multiple entities—such as states, firms, or individuals—the resulting structure is referred to as *time-series cross-sectional data*, or more simply, *panel data*. Panel data can be *balanced*, when all variables are observed over the same time span for every unit, or *unbalanced*, when at least one unit (i.e., grouping factor) has missing observations for at least one time period.

A further, core distinction concerns how the model accounts for unit-specific, time-invariant components (in this study, the U.S. states included in the SIR dataset). In a *fixed-effects* model, each unit is assigned a fixed intercept that captures all characteristics that do not vary over time. In this framework, unobserved heterogeneity across units is not explicitly modeled; in other words, the effects of time-invariant variables—such as structural differences between states—are not estimated. Conversely, a *mixed-effects* (or *random-effects*) model treats unit-specific effects as random variables drawn from a common normal distribution. This allows the model to estimate their variance, thereby capturing how the time-invariant characteristics of one unit relate to those of others. Both approaches can be integrated within a *Hierarchical Linear Modeling* (HLM) framework, where fixed effects represent the average impact of predictors across the entire population, and random effects account for unit-specific deviations from this overall trend (Baltagi, 2013; Chen, 2021).

In our study, we aggregate monthly relative risks across all units to estimate the overall risk of hospitalization and amputation, while still allowing for state-level deviations by computing relative risks relative to each state's population rather than the national total. Although hierarchical modeling provides a more flexible framework, explicitly modeling between-state variation falls beyond the scope of the present research.

The question whether to implement a global as opposed to a local forecasting model has been extensively addressed by Montero-Manso and Hyndman (2021). In their study, they found that global methods show good results not only in groups of similar time series, but even in heterogeneous groups. When a large number of time series is available, global models may generalize better with fewer parameters. Importantly, global models can afford much larger memory in terms of lags, being able to capture long-memory patterns that local models could learn only if manually

engineered. We empirically test these findings by training different models, both linear and nonlinear ones, locally – fitting a single function to each cross-sectional time-series – and globally – by pooling all state-level series into a single learning problem.

Such a global model strategy raises the question of how panel data should be pooled. Panels can be combined on both the temporal as the time-invariant dimension, or along each axis separately. Appropriate handling of the panel structure during the temporal split of the data into a train and a test set can help prevent a wide range of potential data leakage (Cerqua, Letta, & Pinto, 2024).

Finally, hierarchical forecasting methods are designed to produce forecasts at multiple levels of aggregation. In our case, we could forecast total hospitalizations at a national and a state level, and with enough data at a county level. Hierarchical approaches reconcile forecasts so that the sum of the disaggregated predictions equal the aggregate forecast. However, we excluded these methods because the target variables are relative risks, which are not additive across hierarchical levels (Hyndman & Athanasopoulos, 2018).

3.4 *Literature Gaps & Contribution*

From the above review we can conclude that the SIR dataset has been used to gain descriptive insights into the temporal dimension of the injury report as well as to classify injury severity based on multiple categorical features. However, much remains unanswered when it comes to how severe injuries evolve over time and across states from a forecasting perspective. A truly forecasting framing is thus worthy of being developed, as it could improve risk monitoring and enable decision makers on a proactive management of prevention resources. Moreover, we contribute to the discussion in the forecasting community about global versus local modelling of panel data by providing an empirical comparison of both approaches in a novel domain, using a unified feature set and consistent evaluation metrics.

4 METHODS

As discussed in subsection 3.3, a crucial methodological consideration for panel time-series data is whether to fit a single function to all series or to repeatedly fit a function to each series. In panel data, grouping factors are referred to as *cross-sectional units*. In this study, each state is observed over the same time span and shares the same features as the remaining units. Therefore, in the SIR dataset a single time series at a state level is a cross-sectional unit.

In a Global Forecasting Model cross-sectional units are assumed to be generated from the same data-generating process. In a Local Forecasting Model each time series is independent (Hewamalage, Bergmeir, & Bandara, 2022). In the following subsections, we illustrate how we constructed the dataset in a way suitable to both time-series forecasting frameworks.

Another key distinction concerns the forecasting strategy. Traditionally, most researchers compared statistical methods against a one-step ahead forecast horizon, where the forecaster predicts the next step in a time series. Moreover, classical models such as ARIMA and Vector Autoregression (VAR) are able to generate recursive multi-step forecasts natively, thereby lowering the need to target the multi-step task with dedicated forecasting strategies. With the advent of machine learning methods for time series forecasting, researchers must design explicit forecasting strategies, as those models lack the built-in capacity to roll forward recursive predictions. Under the recursive strategy, a model generates a one-step ahead forecast which is fed back to the historical window, and used to recompute all the features, before the model can generate a new one-step ahead forecast. This process is repeated until the final time point in the horizon is reached.

In their seminal review, Taieb, Bontempi, Atiya, and Sorjamaa (2011) formalized five strategies commonly used for multi-step forecasting: (i) recursive, (ii) direct, (iii) direct-recursive hybrids, (iv) multi-input multi-output, and (v) direct multi-output. The most relevant strategy for answering our research questions is the direct strategy. Because our models are built under different assumptions - linear, tree-based, and artificial networks - the direct approach allows for comparability by deploying the same pipeline under the same evaluation paradigm.

In fact, under the direct approach, a separate model is independently trained to predict a single output for each horizon. For instance, forecasting four horizons (e.g., $h = 1, 3, 6, 12$) requires fitting four separate models, one per horizon. Unlike the recursive strategy, the direct approach does not use forecasts as inputs in the historical window. As a consequence, errors do not compound at each next step, overcoming the risk of bias accumulation over longer horizons.

As a further motivation, the data complexity of the SIR data set is estimated to be sufficiently low to run separate models at a reasonable computation time. One of the drawbacks of the direct approach is the computational burden, since a separate model has to be fitted for every forecast horizon. For RQ1 and RQ2, the models are evaluated on a direct one-step-ahead forecast over the entire train set (12 months) to ensure methodological consistency.

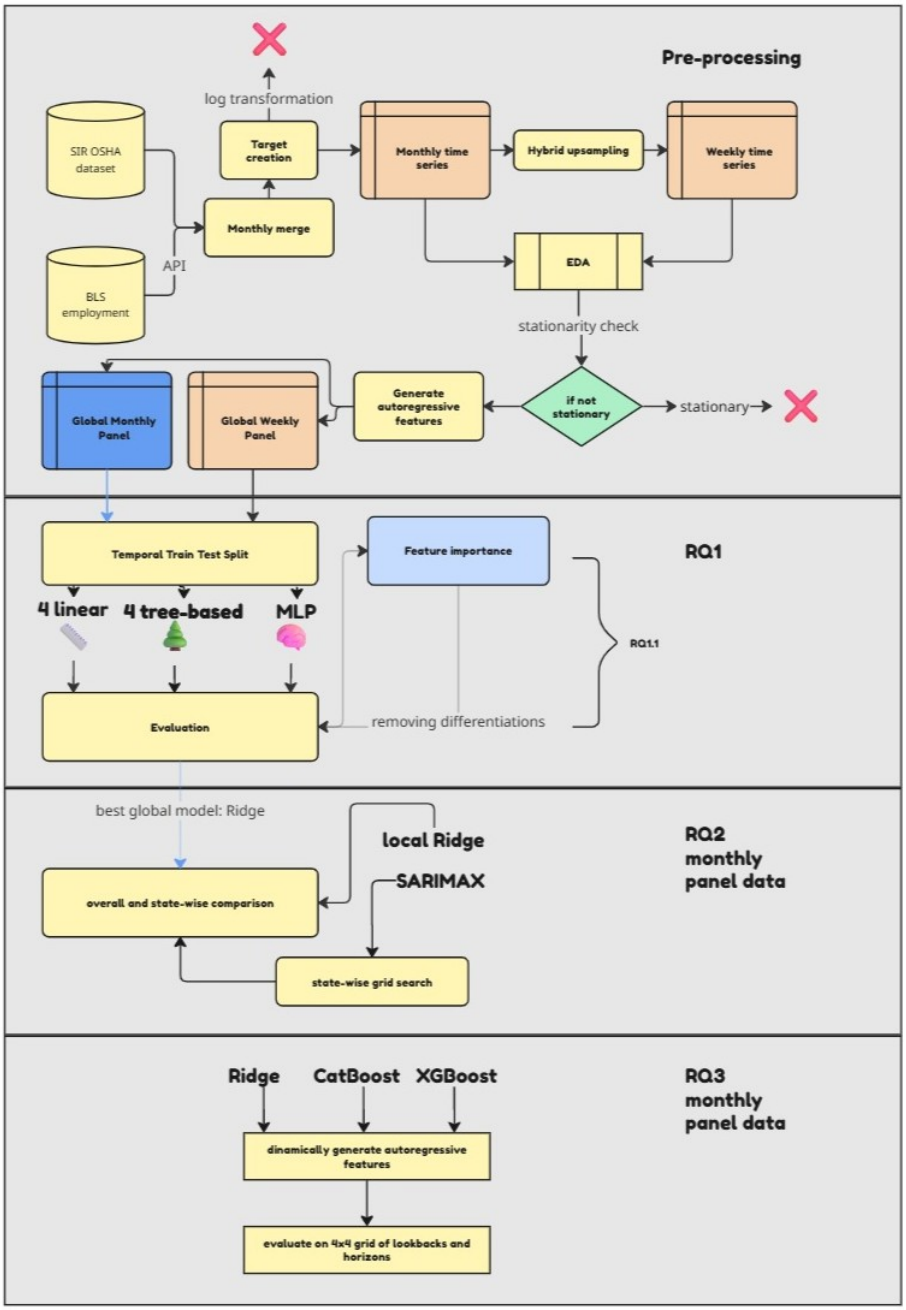

Figure 1: Workflow

4.1 *Dataset Description*

The dataset is publicly available through the OSHA Severe Injury Dashbord [1]. OSHA collects reports about severe injuries occurred within U.S. states under its jurisdiction. There reports form the Severe Injury Reports (SIR) dataset. There are 34 states with a federal plan under OSHA jurisdiction, although the SIR dataset also contains reports from non-federal-plan states when the injured employees are federal workers. In addition, seven states have state plans covering only state and local government workers, while private-sector workers are covered by federal plans. To ensure a fair comparison between states, we included only states operating under a federal OSHA jurisdiction, together with the seven states mentioned above, as all reports from those states are collected from the private sector.

The dataset contains 98,801 entries, spanning from January 1, 2015 to February 28, 2025. Each entry represents a report submitted by an employer. A single report may mention multiple severe injuries (e.g. one hospitalization and one amputation), as employers are required to specify *"the number of employees who suffered a fatality, in-patient hospitalization, amputation, or loss of an eye"*[2]. This means that the features used to construct the target series (Hospitalization and Amputation) are count variables, not binary encodings, and that we can not aggregate the data at the report level. The dataset provides 26 attributes, ranging from geographical information about the city, the address, and the state where the accident occurred, to details about the nature of the injury, the body part affected, the employer, and the business sector. Only variables required to construct state-level monthly injury counts and relative risks were retained.

To make our model more transparent for decision makers, we derived two new target variables from the two attributes Hospitalization and Amputation. Since these two attributes represent only absolute numbers, they may vary depending on the worker population. More populated states will probably report a larger number of severe injuries than states with a small workforce. Even within the same state, workforce trends over time can enormously impact the number of accidents. As an example, the COVID period saw a dramatic shrinking of the workforce, possibly leading to a number of severe injuries much lower than in any other period.

To achieve interstate comparability, data from the Bureau of Labor Statistics (BLS) were fetched about seasonally adjusted, monthly employment at a state level. Employment is estimated on nonfarm, payroll jobs by the Current Employment Statistics (CES) survey. The covered time period

[1] https://www.osha.gov/severe-injury-reports. The dataset was retrieved on August 28, 2025.

[2] https://www.osha.gov/laws-regs/regulations/standardnumber/1904/1904.39

spans from January 1, 2015 to December 31, 2024. Three states, namely American Samoa, Guam, and Northern Mariana Islands, were excluded from the panel since they are not covered by the BLS. After inspecting the testing set, one more U.S. territory, Virgin Islands, was excluded because of few data points. The ratio of both target features over the monthly employment was separatedly computed for each state. Therefore, the target to be predicted is not the absolute count of injuries per state, but the relative risk of getting either hospitalized or amputated per 100,000 state employees.

4.2 *Data Cleaning and Preparation*

The main preparation step involved aggregating both hospitalization and amputation risks at the monthly level to ensure a consistent merge with the monthly state employment statistics from the Bureau of Labor Statistics (BLS). After merging the two datasets, some missing values originated from a mismatch between the SIR time period (ending in February 2025) and the BLS time period (ending in December 2024). These entries were removed from the panel.

For the weekly forecasting task, the state employment data were upsampled from monthly to weekly frequency using a hybrid strategy. Missing weekly values were interpolated with the exception of the period from March to September 2020. During this interval, the COVID-19 pandemic caused an abrupt drop in employment that the interpolation method was not able to capture properly. For this particular period, each initial monthly value was forward-filled until the next observed monthly value. Upsampling monthly employment data to weekly frequency assumes relatively smooth workforce trends within each month.

4.3 *Exploratory Data Analysis*

A skewed data distribution of the target feature may lead to poor generalization, as the time series can be sensitive to extreme values.

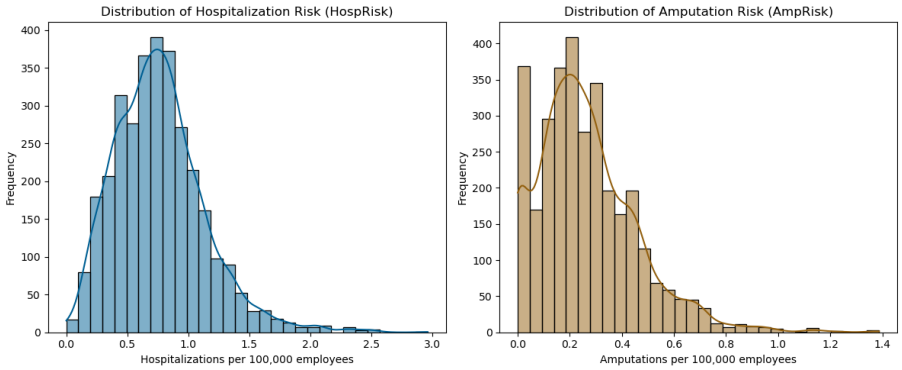

Figure 2: Distributions of targets

Since both hospitalization and amputation risk show a right-skewed distribution, we initially standardized these features using a log transformation. To test whether this transformation improved predictive accuracy, we trained a Ridge regressor on the log-transformed targets and compared the performance to the model trained on targets on the original scale. As the accuracy deteriorated, it is likely that the transformation removed structural components that the regressor could have exploited. Since we normalized the targets by adjusting for the monthly employment population for each state, further standardization offered no additional benefit.

We also experimented with differencing the global mean time series – monthly aggregation across all states – before generating lagged features. In fact, after conducting an Augmented Dickey-Fuller test on each state-level series to check for stationarity, we found that some series are stationary, while others are not. Similar to the log transformation, the differentiation increased both RMSE and MAE on the one-step-ahead twelve-month forecast. Since modern machine-learning models do not strictly require stationarity assumptions, we preferred to inform the models using lagged values, rolling means, and differentiations of lagged values on the original scale.

4.4 *Feature Engineering*

In this crucial step, each time series was provided with exogenous variables. Although it is not strictly necessary for a ML model to be explicitly informed about data patterns, exogenous variables can still benefit the model's capabilities to predict temporal values. This is a viable option when the data complexity is low, the data are likely to be properly fitted by a linear model, and seasonal patterns can be detected from a visual

inspection of the time series. The exogenous variables are divided into four categories:

Table 1: Overview of target variables, autoregressive features, and exogenous features used in the monthly forecasting models.

Category	**Variables**
Target variables	`HospRisk`: monthly hospitalization risk per 100,000 employees; `AmpRisk`: monthly amputation risk per 100,000 employees.
Calendar and seasonal features	`Year`, `Month`, `Quarter`, `DaysInMonth`; Fourier terms: `Month_sin`, `Month_cos`.
Lagged features	`HospRisk_lag1`, `HospRisk_lag3`, `HospRisk_lag6`, `HospRisk_lag12`; `AmpRisk_lag1`, `AmpRisk_lag3`, `AmpRisk_lag6`, `AmpRisk_lag12`.
Differenced features	First differences: `HospRisk_diff1`, `AmpRisk_diff1`; seasonal (12-month) differences: `HospRisk_diff12`, `AmpRisk_diff12`.
Rolling features	3-month rolling means: `HospRisk_roll3`, `AmpRisk_roll3`; 6-month rolling means: `HospRisk_roll6`, `AmpRisk_roll6`.
Exogenous variables	`State` (categorical identifier, one-hot encoded in the models)

5 RESULTS

This section presents the results from our experiments, following the research questions order. We will proceed by displaying the results for Hospitalization Risk, first including all the autoregressive features and second, to answer sub-question 1.2, without the differentiations. We will compare monthly and weekly results.

Forecasts for the monthly global model are based on 2859 train points and 358 test points, whereas forecasts for the weekly global model are based on 10658 train points and 1354 test points.

Unfortunately, during the construction of the autoregressive features for framing the time series as a supervised regression task, we incorrectly define our model in a way that invalidates the forecasting results reported in this section. In the current model the feature matrix included both lagged values (e.g., y_{t-1}, y_{t-12}) and differentations (e.g., Δy_t, $\Delta_{12} y_t$). Because differenced values are functions of the original lags, the model was able to see the target variable. This created perfect multicollinearity in the design matrix (VIF $\to \infty$), leading to unrealistically low RMSE values. We will

revise how to construct the autogressive feature to prevent any other data leakage.

5.1 *Hospitalization risk at monthly resolution with all autoregressive features*

The performance of the candidate models was tested on a one-step-ahead twelve-month forecast horizon, using the same autoregressive matrix. As we do not have a benchmark from other studies, we compared the results against a persistence model, where the last observation in a time series is carried forward to predict all future timesteps.

Monthly Hospitalization Risk

Model	**RMSE$_{train}$**	**MAE$_{train}$**	**MASE$_{train}$**	**RMSE$_{test}$**	**MAE$_{test}$**	**MASE$_{test}$**
Ridge	0.00158	0.00113	0.00464	**0.00143**	**0.00102**	**0.00417**
CatBoost	0.02818	0.02056	0.08448	**0.03000**	**0.02155**	**0.08854**
XGBoost	0.02191	0.01674	0.06879	**0.03138**	**0.02312**	**0.09499**
LightGBM	0.02883	0.01966	0.08076	**0.03319**	**0.02509**	**0.10307**
MLP	0.04278	0.03373	0.13857	**0.03954**	**0.03191**	**0.13109**
PLS	0.13683	0.09853	0.40480	**0.12267**	**0.08850**	**0.36360**
AdaBoost	0.15088	0.12198	0.50115	**0.16247**	**0.13025**	**0.53512**
Lasso	0.37658	0.28726	1.18020	**0.34740**	**0.27665**	**1.13663**
ElasticNet	0.37658	0.28726	1.18020	**0.34740**	**0.27665**	**1.13663**
Naive persistence	–	–	–	**0.34967**	**0.24340**	**1.00000**

Table 2: Train and test performance metrics for all evaluated models on *monthly hospitalization risk*. Test-set metrics are highlighted in bold.

The Ridge regressor shows consistent and pronounced advantage on all the other models, including the linear ones Lasso and ElasticNet. The Rdige prediction errors are an order of magnitude smaller than those of the next-best model, CatBoost. Tree-based models perform similarly, with test RMSE values between 0.030 and 0.033.

A linear model such as Ridge tends to perform optimally on highly collinear features within a smooth time series. By inspecting the correlation among a subset of autoregressive features, we see that a first order differentiation is negatively correlated with a lag of one. A ridge regressor shrinks the coefficients by adding a weight decay regularization, which is defined as L2 norm. This forces the low-informative features to approximate zero, acting as a sort of feature selection.

In our case, it is evident that the ridge regressor relied on a small subset of highly correlated features, namely the differentiations and the lags, thereby capturing the trend of each single time series in the panel data. Conversely, the regularized linear regressors Lasso and ElasticNet,

without any hyperparameter tuning, applied a too aggressive L1 weight decay, which can force coefficients to be zero, thereby completely removing the autoregressive features from the input matrix.

Tree-based models, by contrast, do not benefit from a small, highly-linear feature space. Their ability to capture nonlinearity is not effective where correlated autoregressive features reduce the diversity of tree splits. This is further confirmed by the low performance of a neural network, the Multi Layer Perceptron, which is best suited for a larger parameter space and greater variability.

To evaluate how well the models generalize on the hold-out test set, we compared the RMSE on the train set against the test set. Whereas the ridge model further benefited from the test set observations, showing an effective way to generalize beyond the train set, tree-based models show some signs of overfitting. While the MLP achieved a slightly lower RMSE, its performance is still below that of the persistence model.

Weekly Hospitalization Risk

Model	**RMSE$_{train}$**	**MAE$_{train}$**	**MASE$_{train}$**	**RMSE$_{test}$**	**MAE$_{test}$**	**MASE$_{test}$**
Ridge	0.00097	0.00068	0.00661	**0.00084**	**0.00059**	**0.00574**
CatBoost	0.01090	0.00708	0.06871	**0.00969**	**0.00649**	**0.06301**
XGBoost	0.00957	0.00677	0.06570	**0.01024**	**0.00691**	**0.06713**
LightGBM	0.01140	0.00703	0.06828	**0.01033**	**0.00672**	**0.06521**
MLP	0.03016	0.02604	0.25281	**0.02986**	**0.02577**	**0.25022**
PLS	0.05691	0.04037	0.39201	**0.05024**	**0.03560**	**0.34571**
AdaBoost	0.06530	0.05256	0.51035	**0.06994**	**0.05682**	**0.55173**
Lasso	0.13654	0.09688	0.94071	**0.12093**	**0.09264**	**0.89953**
ElasticNet	0.13654	0.09688	0.94071	**0.12093**	**0.09264**	**0.89953**
Naive persistence	–	–	–	**0.14694**	**0.10298**	**1.00000**

Table 3: Train and test performance metrics for all evaluated models on *weekly hospitalization risk*. Test-set metrics are highlighted in bold.

6 DISCUSSION

6.1 *Answering Research Questions*

In this section, the results are placed in the broader context of occupational risk prevention. For each research question, we summarize our findings and assess whether they align with prior work. Since empirical benchmarks for forecasting occupational injury risk are scarce, we will primarly situate our findings within the theoretical and methodological literature on global versus local modelling strategies in panel time series forecasting.

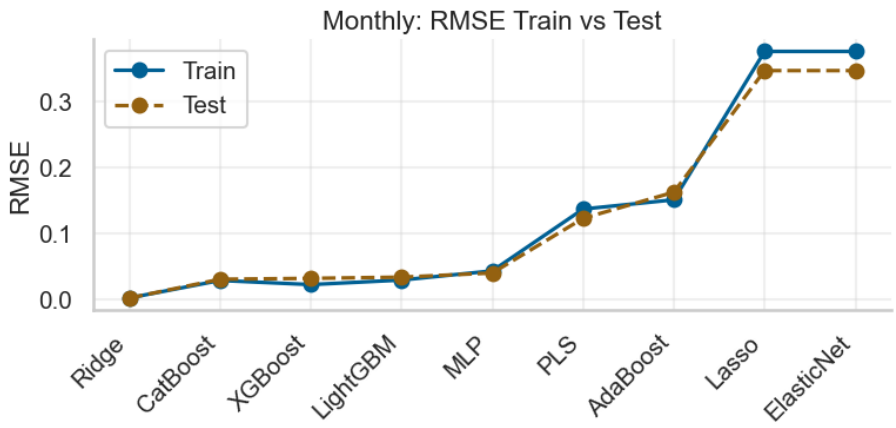

Figure 3: Monthly RMSE: Train vs Test

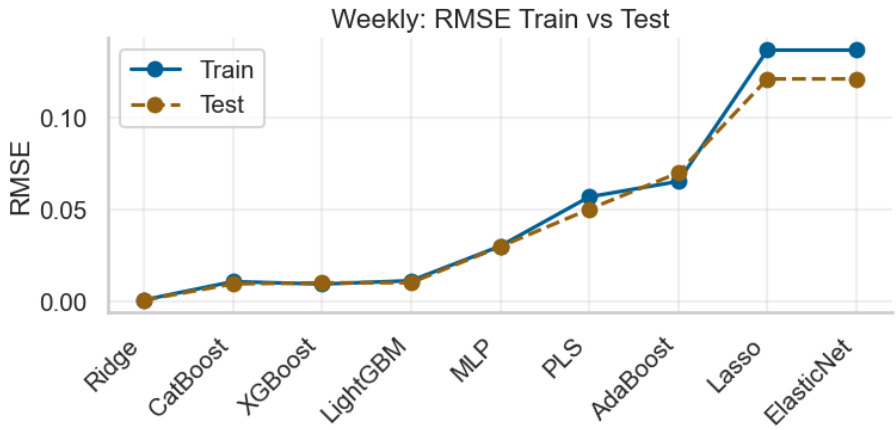

Figure 4: Monthly RMSE: Train vs Test

To answer research question 1, we evaluated nine different models belonging to three family models: linear ones (ridge, lasso, elasticnet, and PLS); tree-based ones (CatBoost, XGBoost, AdaBoost, and LightGBM), and one neural network (MLP). Our findings confirm that a simple feature space can be better modelled by linear regressors. While most studies focus on classification tasks with diverse subsets of input features, our problem was simplified by having two target variables and a feature matrix mostly composed of autoregressive features. This simple set up showed that a ridge regressor can efficiently capture most of the variation in the time series by leveraging on long history data even when aggregated at a monthly level. However, when the ridge regressor is fitted locally, its accuracy does not significantly differ from the tree-based models fitted on the global forecasting panel. This is the key finding in our study, which

aligns with previous research: by pooling panel data the models can better capture the time series than by fitting a single model to each time series.

Among the tree-based models, AdaBoost has the higher RMSE, a finding that is consisten with the most recent literature. Once again, it is important to stress that our problem is intrinsically different, because it is a regression problem. Given the nature of the input data, we might have expected a SARIMAX model to performcompetitevely. However, the local strategy does not enable the model to customize its forecasts to each state time series.

6.2 *Limitations and further research*

7 CONCLUSION

REFERENCES

Baltagi, B. H. (2013). Chapter 18 - Panel Data Forecasting. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 995–1024). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B978044462731500018X (ISSN: 1574-0706) doi: https://doi.org/10.1016/B978-0-444-62731-5.00018-X

Bojer, C. S. (2022, October). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, *38*(4), 1555–1561. Retrieved 2025-11-17, from https://linkinghub.elsevier.com/retrieve/pii/S0169207021001771 doi: 10.1016/j.ijforecast.2021.11.003

Cerqua, A., Giannantoni, C., Letta, M., & Pinto, G. (2024, December). *'Dead Man Working': A Place-based Approach to Workplace Fatalities* [SSRN Scholarly Paper]. Rochester, NY: Social Science Research Network. Retrieved 2025-09-08, from https://papers.ssrn.com/abstract=5040905 doi: 10.2139/ssrn.5040905

Cerqua, A., Letta, M., & Pinto, G. (2024). *On the (Mis)Use of Machine Learning with Panel Data.* SSRN. Retrieved 2025-11-07, from https://www.ssrn.com/abstract=5014594 doi: 10.2139/ssrn.5014594

Chen, J. M. (2021, February). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, *27*(1), 1–16. Retrieved 2025-11-07, from https://doi.org/10.1007/s11294-021-09815-6 doi: 10.1007/s11294-021-09815-6

Gomes, H., Parasram, V., Collins, J., & Socias-Morales, C. (2023, September). Time series, seasonality and trend evaluation of 7FIX ME!!!!years (2015–2021) of OSHA severe injury data. *Journal of Safety Research*, *86*,

30–38. Retrieved 2025-09-14, from https://www.sciencedirect.com/science/article/pii/S0022437523000798 doi: 10.1016/j.jsr.2023.06.005

Hall, T., & Rasheed, K. (2025, January). A Survey of Machine Learning Methods for Time Series Prediction. *Applied Sciences*, *15*(11), 5957. Retrieved 2025-11-17, from https://www.mdpi.com/2076-3417/15/11/5957 (Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/app15115957

Han, C., Zhou, Y., Sun, J., & Li, Z. (2025, November). An optimized ridge regression for forecasting time series with a fixed period. *Pattern Recognition Letters*, *197*, 274–281. Retrieved 2025-12-22, from https://www.sciencedirect.com/science/article/pii/S0167865525002971 doi: 10.1016/j.patrec.2025.08.017

Hasan Khalleefah Hassan, M., & Khalifa, W. M. S. (2025). Work Place Safety: Machine Learning Techniques for Assessing Workplace Incident Severity. *IEEE Access*, *13*, 34211–34226. Retrieved 2025-09-10, from https://ieeexplore.ieee.org/abstract/document/10891569 doi: 10.1109/ACCESS.2025.3543136

Hewamalage, H., Bergmeir, C., & Bandara, K. (2022, April). Global models for time series forecasting: A Simulation study. *Pattern Recognition*, *124*, 108441. Retrieved 2025-11-07, from https://www.sciencedirect.com/science/article/pii/S0031320321006178 doi: 10.1016/j.patcog.2021.108441

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. Retrieved 2025-11-29, from https://research.monash.edu/en/publications/forecasting-principles-and-practice-2/ (Publisher: OTexts)

Johnson, M. S., Levine, D. I., & Toffel, M. W. (2023, October). Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA. *American Economic Journal: Applied Economics*, *15*(4), 30–67. Retrieved 2025-11-06, from https://pubs.aeaweb.org/doi/10.1257/app.20200659 doi: 10.1257/app.20200659

Khairuddin, M. Z. F., Lu Hui, P., Hasikin, K., Abd Razak, N. A., Lai, K. W., Mohd Saudi, A. S., & Ibrahim, S. S. (2022, October). Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *International Journal of Environmental Research and Public Health*, *19*(21), 13962. Retrieved 2025-09-08, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9653932/ doi: 10.3390/ijerph192113962

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023, August). A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Net-

works. *Future Internet*, *15*(8), 255. Retrieved 2025-11-29, from https://www.mdpi.com/1999-5903/15/8/255 (Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/fi15080255

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022, October). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. Retrieved 2025-11-29, from https://www.sciencedirect.com/science/article/pii/S0169207021001874 doi: 10.1016/j.ijforecast.2021.11.013

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023, February). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. Retrieved 2025-11-07, from https://onlinelibrary.wiley.com/doi/10.1111/joes.12429 doi: 10.1111/joes.12429

Michaels, D., & Wagner, G. R. (2025, April). OSHA Injury Data: An Opportunity for Improving Work Injury Prevention. *American Journal of Public Health*, *115*(4), 588–595. Retrieved 2025-09-04, from https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2024.307934 (Publisher: American Public Health Association) doi: 10.2105/AJPH.2024.307934

Montero-Manso, P., & Hyndman, R. J. (2021, March). *Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality*. arXiv. Retrieved 2025-11-05, from http://arxiv.org/abs/2008.00444 (arXiv:2008.00444 [cs]) doi: 10.48550/arXiv.2008.00444

Organization, I. L. (2020). *Safety + health for all : an ILO Flagship Programme : key facts and figures (2016-2020) - International Labour Organization.* Retrieved 2025-11-29, from https://labordoc.ilo.org/discovery/fulldisplay/alma995108293102676/41ILO_INST:41ILO_V2

Qu, R., Timmermann, A., & Zhu, Y. (2024, July). Comparing forecasting performance with panel data. *International Journal of Forecasting*, *40*(3), 918–941. Retrieved 2025-11-08, from https://www.sciencedirect.com/science/article/pii/S0169207023000766 doi: 10.1016/j.ijforecast.2023.08.001

Rahman, M. M., Hossain, A., & Sikder, M. A. (2024, May). Machine Learning Applications in Industry Safety: Analysis and Prediction of Industrial Accidents. In *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)* (pp. 1–6). Retrieved 2025-11-07, from https://ieeexplore.ieee.org/document/10561314 doi: 10.1109/ICSSES62373.2024.10561314

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., ... Shaman, J. (2019, February). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of*

the United States of America, *116*(8), 3146–3154. Retrieved 2025-11-29, from https://pmc.ncbi.nlm.nih.gov/articles/PMC6386665/ doi: 10.1073/pnas.1812594116

Rostami-Tabar, B., Goltsos, T. E., & Wang, S. (2023, February). Forecasting for lead-time period by temporal aggregation: Whether to combine and how. *Computers in Industry*, *145*, 103803. Retrieved 2025-12-20, from https://www.sciencedirect.com/science/article/pii/S0166361522001993 doi: 10.1016/j.compind.2022.103803

Taieb, S. B., Bontempi, G., Atiya, A., & Sorjamaa, A. (2011, August). *A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition.* arXiv. Retrieved 2025-11-13, from http://arxiv.org/abs/1108.3259 (arXiv:1108.3259 [stat]) doi: 10.48550/arXiv.1108.3259

Vivian, G. A., Bauder, R. A., & Khoshgoftaar, T. M. (2025, July). A comprehensive survey on machine learning for workplace injury analysis: risk prediction, return to work strategies, and demographic insights. *Journal of Big Data*, *12*(1), 167. Retrieved 2025-09-01, from https://doi.org/10.1186/s40537-025-01229-z doi: 10.1186/s40537-025-01229-z

APPENDIX A

Appendix A (page 21).