Forecasting Weekly Depression Incidence in Mexico: A Multi-Model Framework with Auditable Per-Series Model Selection

Javier A. Rebull-Saucedo[1](✉), Juan Carlos Pérez-Nava[2], Luis Gerardo Sánchez-Salazar[3], Grettel Barceló-Alonso[4], and Ruth Manuela Pérez-Hernández[2]

[1] Santander Bank US, Massachusetts, USA
rebull@outlook.com
[2] Mexican Social Security Institute (IMSS), Mexico City, Mexico
[3] Tesla, Inc., California, USA
[4] Tecnológico de Monterrey (ITESM-HGO), Hidalgo, Mexico

Abstract. Weekly forecasting of psychiatric demand is hard because surveillance signals are heterogeneous across regions, sexes and time, so no single forecasting paradigm performs uniformly well. Rather than impose one algorithm across an entire surveillance system, we present a deterministic, auditable framework that selects, for each series, the model best supported by its own evidence. The framework draws on a library of forecasters with deliberately different inductive assumptions—a structural additive decomposition, a probabilistic recurrent network and two tree-based hybrids—so the candidate pool spans diverse temporal representations and improves coverage across this heterogeneity. A transparent rule then picks one model per series on cross-validated accuracy, reassigns low-incidence series to a regional model, and records every decision in a justified selection log. We instantiate it for weekly incidence of depression (CIE-10 F32) across the 32 Mexican federal entities, drawing on the public bulletins of the National Epidemiological Surveillance System (SINAVE), and validate it twice: by rolling-origin cross-validation over a decade of weekly observations and by a prospective out-of-sample assessment against the 2026 bulletins released after the training cut-off. The resulting assignment concentrates on the recurrent model for most federal-entity series while retaining simpler models where aggregation favours them, and the separately locked national forecast tracks the early-2026 bulletins. Because cross-validated accuracy need not persist as the live signal shifts, the selection is revisable—it re-selects as new bulletins arrive—so the system stays auditable and accountable in deployment, not a frozen leaderboard winner. It is reproducible and extends to other conditions with heterogeneous behaviour.

Keywords: Depression · Time series forecasting · Probabilistic recurrent networks · Multi-model forecasting · Epidemiological surveillance · Model selection.

1 Introduction

Depression is a leading cause of disability worldwide and a major burden on public-health systems [25]. In Mexico it affects an estimated 8–10 million people, and the National Epidemiological Surveillance System (SINAVE) records over 150,000 new cases of depressive episode (CIE-10 F32) per year [11]. Despite the routine availability of weekly case counts, the planning of psychiatric services remains largely *reactive*: capacity is adjusted only after demand surges are already visible in the bulletin, producing recurrent saturation windows during the February–May period.

Accurate weekly, federal-entity forecasts would make this planning *proactive*: a reliable 52-week-ahead trajectory lets authorities pre-position staff ahead of seasonal peaks, size programmes per state, and time help-seeking campaigns for the quieter months. The problem is hard precisely where it matters. The depression signal is strongly heterogeneous: case loads differ by orders of magnitude across entities, are sharply skewed by sex, and range from smooth metropolitan series to volatile low-volume states, over a panel marked by a COVID-19 regime change and a post-pandemic level shift. No single paradigm accommodates all of these at once: structural decompositions excel on smooth aggregates but lag on noisy series, whereas representation-learning models share structure across series but can over-react on smooth aggregates. Deployment also raises a trust problem: a model chosen on historical cross-validation can quietly lose its edge once the live signal shifts, so an institution needs a selection it can audit and revise, not a one-off leaderboard winner.

We address this with a **multi-model forecasting framework** (referred to throughout as *the proposed framework*). Rather than committing to one model family, it trains four forecasters with complementary inductive assumptions for each series and selects one per series through a deterministic, auditable rule, so that whichever paradigm fits a given series best is the one that produces its forecast. Although evaluated here on depression, the framework is designed for epidemiological surveillance in general: choosing one model per series is a direct response to the heterogeneity that makes any single architecture unlikely to dominate across the many conditions a surveillance system monitors.

The main contributions of this study are threefold: (i) we present and empirically study a multi-model forecasting framework for psychiatric surveillance that places structural, recurrent and tree-based forecasters behind a single interface, enabling a head-to-head, per-series comparison under one rolling-origin protocol; (ii) we introduce a deterministic, auditable per-series model-selection strategy that ranks candidates on cross-validated symmetric error with scale-free and large-error tiebreakers and a transparent low-incidence regional fallback, producing a fully justified, revisable selection log—an auditable selection mechanism built for trustworthy deployment under distribution shift, not a one-off leaderboard choice; and (iii) we provide a nation-wide empirical characterisation, over 111 series–stratum combinations and 626 weeks of SINAVE data, of the regional, demographic and temporal heterogeneity of the depression signal and of the conditions under which each forecasting paradigm prevails. A central finding

is that the cross-validated ranking is a *selection* signal, not an out-of-sample guarantee: per-series accuracy degrades on live 2026 data, which the revisable rule absorbs by reassigning models as the signal accumulates—substantiated on the first eighteen epidemiological weeks of 2026.

2 Foundations and Related Work

2.1 Epidemiological Time-Series Forecasting

Incidence forecasting has progressed from linear, series-by-series statistical models to global, representation-learning architectures that share parameters across many related series. The Box–Jenkins autoregressive integrated moving average (ARIMA) family [5] remains a standard baseline, but its linearity and stationarity assumptions limit performance on non-stationary public-health signals subject to regime changes. Exponential-smoothing state-space models (ETS) [16] offer a flexible additive or multiplicative decomposition but treat each series independently. Prophet [26] popularised an additive decomposition with explicit trend, seasonality and holiday components and is robust to missing data and outliers, making it a frequent default in public-health pipelines; its smooth additive structure, however, can under-fit the volatile low-count series typical of small federal entities. Recurrent neural networks (RNNs), and in particular long short-term memory (LSTM) networks, have shown clear advantages on epidemiological series: Chimmula and Zhang [8] applied LSTMs to COVID-19 incidence in Canada, outperforming classical baselines on short horizons, though such data-hungry models can over-react on the smooth, high-volume aggregates where simpler decompositions already suffice. For public-health surveillance specifically, collaborative forecasting efforts have shown that systematic, reproducible evaluation of probabilistic forecasts is feasible at national scale [23,9]. More recent architectures push long-horizon accuracy further—for example, the hierarchical-interpolation network N-HiTS [6]—and large pretrained *foundation* models for time series, such as Chronos [3], forecast across many domains with little or no task-specific training; their behaviour on the sparse, regime-shifting panels typical of epidemiological surveillance is still largely untested.

2.2 Foundations of Multi-Model Forecasting

The case for combining heterogeneous forecasters is well established. Salinas et al. [24] introduced DeepAR, a global probabilistic autoregressive RNN that performs *cross-learning* across a panel of related series and emits a full predictive distribution at each step; recent surveys of deep learning for forecasting situate such global models within a broad design space [4,19,14]. Large-scale empirical evidence from the M4 and M5 forecasting competitions—open benchmarks spanning, respectively, 100,000 cross-domain series and a large hierarchical retail-sales corpus—shows that machine-learning and hybrid methods are competitive with, and often superior to, classical univariate methods across heterogeneous

collections, and that combinations are particularly robust [20,21], including automated ensembles tuned for weekly series [13]. Combining the explicit components of additive decompositions with the non-linear capacity of gradient-boosting decision trees (GBDT) has proven effective in retail and energy forecasting [17], and Prophet–neural hybrids have improved pandemic incidence forecasts [10]. Stacked generalisation [28] formalises learning a meta-model on top of heterogeneous base learners. These foundations motivate a design in which structural, statistical and learning-based assumptions coexist and a principled rule decides between them per series.

2.3 Depression and Mental-Health Surveillance

The burden of depression in Mexico has been quantified through the Global Burden of Disease study [12], and the COVID-19 pandemic produced a measurable global increase in depressive-disorder burden [25]. Sex-related disparities in reported incidence are well established clinically [1]. No prior work has applied a multi-model forecasting framework with auditable per-series model selection to the SINAVE depression panel at federal-entity granularity, together with an out-of-sample prospective evaluation.

3 Materials and Methods

3.1 Data Characterisation

The study draws on weekly case counts of depressive episode (CIE-10 code F32, the Spanish-language denomination of the WHO ICD-10 classification used by SINAVE) reported in the SINAVE bulletin by the Dirección General de Epidemiología [11], the consolidated public-health surveillance system of Mexico. After calendar alignment under the SINAVE epidemiological-week convention,[5] the harmonised panel contains **626 weekly observations per series**, spanning 2014-W02 through 2025-W53. Epidemiological week 1 of 2014 was unavailable in the retrieved archive, so 2014 contributes 52 weeks; 2020 and 2025 each contain 53 epidemiological weeks, giving $626 = 10{\times}52 + 2{\times}53$. Records are disaggregated by sex (women, men) and aggregated to a non-disaggregated total (general). We analyse the 32 federal entities, four socio-urban regions defined by population density and INEGI demographic indicators (*Metropolitana alta*, *Urbana media*, *Sur-Sureste vulnerable*, *Rural / dispersa*), and a national series, yielding $37 \times 3 = 111$ series–stratum combinations.

Temporal and seasonal structure. Figure 1 shows the national series. Three regimes are visible: a stable 2014–2018 baseline; the COVID-19 disruption of 2020–2021, when both reporting capacity and clinical demand shifted abruptly;

[5] The SINAVE calendar is similar but not identical to the ISO 8601 and CDC Morbidity and Mortality Weekly Report (MMWR) week conventions. We use the calendar as published in the bulletin and do not re-index against either external standard.

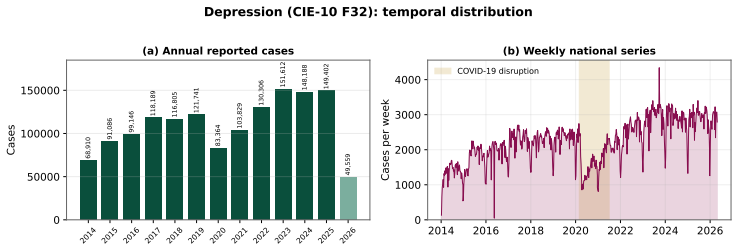

Fig. 1. Depression (CIE-10 F32) temporal distribution. (a) Annual reported cases 2014–2026 (2026 is year-to-date). (b) Weekly national series; the COVID-19 disruption window is highlighted in gold. The post-2022 level shift and the recurring February–May peaks motivate models that adapt across regimes.

and a post-pandemic regime in 2022–2025 that exceeds pre-pandemic levels. The series exhibits clear intra-annual seasonality, with recurring annual peaks between epidemiological weeks 8–20 (February–May).

Spatial, regional and demographic heterogeneity. State-level case loads vary by more than two orders of magnitude (Fig. 2): the highest-volume entities (Mexico City, the State of Mexico, Jalisco, Chihuahua, Veracruz) together account for about 40% of the cumulative caseload, while sparsely populated southern states are lower but more volatile; the *Metropolitana alta* region is the most stable and *Sur-Sureste vulnerable* the most variable relative to its mean, which penalises single-series baselines. Women account for 74.0% of reported cases against 26.0% in men over 2014–2025, exceeding the 1.5–2× female prevalence of major depressive disorder [1]; as the data are *reported incidence*, this likely reflects both clinical patterns and male under-reporting. We forecast the three strata separately and revisit the gap in Sect. 6. By the indicators we can measure (median annual coefficient of variation below 0.25, negligible missing weeks, a well-contained COVID-19 window), this panel is unusually clean, making it amenable to data-driven sequence models.

3.2 Proposed Framework

Rationale. The heterogeneity documented above motivates a design that does not commit to one inductive bias. We train, for each series, four forecasters whose assumptions are deliberately complementary, and we let a deterministic rule choose between them. The methodological core is not any individual forecaster—all four are established models—but the principled, per-series choice among them: each series is matched to the inductive bias its own history supports, through a deterministic and fully auditable rule rather than a global or hand-tuned commitment. A structural additive model captures explicit trend and seasonality and is strong on smooth aggregates; a probabilistic recurrent network with a flexible autoregressive structure is strong on the many noisy

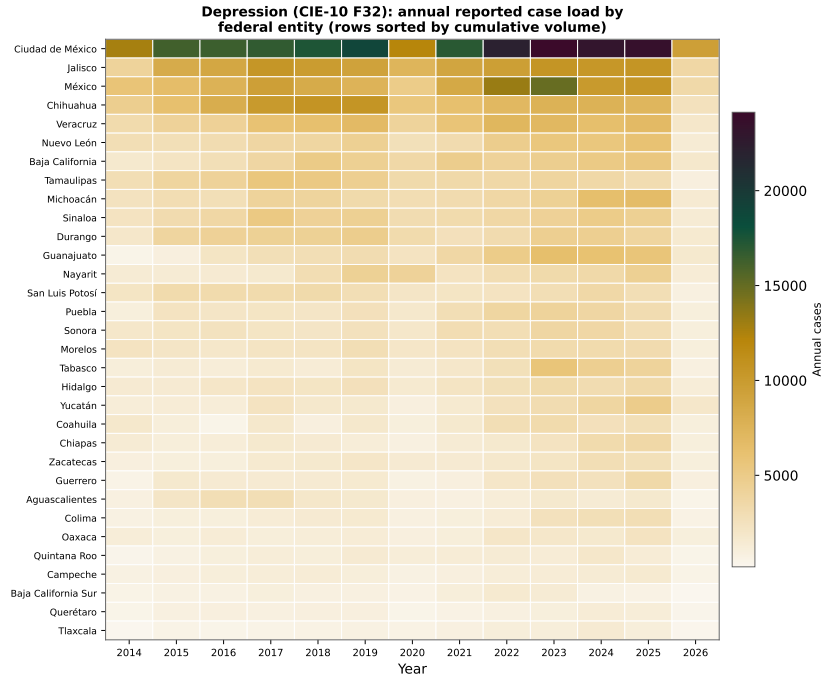

Fig. 2. Annual reported depression (CIE-10 F32) case load by federal entity, 2014–2026 (2026 year-to-date), with rows sorted by cumulative volume. Annual loads span more than two orders of magnitude across entities—from high-volume states such as Mexico City, Jalisco and the State of Mexico down to the sparsely populated south—and the five highest-volume entities account for about 40% of the cumulative caseload. This per-entity spread is the heterogeneity the proposed framework is designed to absorb through per-series model selection.

state series; two tree-based hybrids add non-linear interactions between calendar, lag and demographic features. The four forecasters are exposed through a single, common interface, so that the selection rule operates uniformly over all of them (Fig. 3).

– **Probabilistic recurrent network (DeepAR)** [24,2]: an autoregressive recurrent network that emits a Student-t distribution whose mean is the point forecast. Each federal-entity and regional series is fitted with its own DeepAR model; the national aggregate uses a single DeepAR trained jointly over the 32 state series, with the per-state forecasts summed to the national total. A controlled comparison (Sect. 3.3) shows that joint training adds only a small margin over per-series fitting, so the per-series variant is used at the federal-entity level.

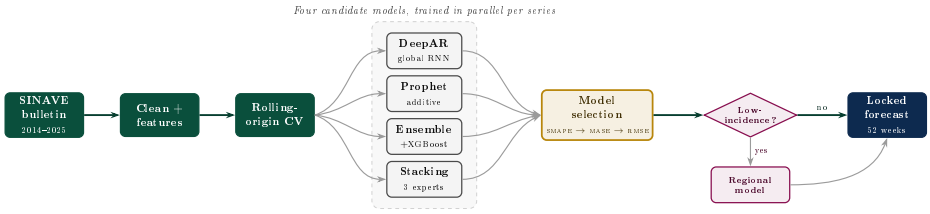

Fig. 3. Proposed framework. Four candidate models with complementary inductive assumptions are trained in parallel per series; a deterministic rule selects one on cross-validated SMAPE (with MASE and RMSE tiebreakers), after which structurally low-incidence series are routed to their regional model before the 52-week forecast is locked.

- **Structural additive model (Prophet)** [26]: piecewise-linear trend with automatic change points plus Fourier annual seasonality, fitted to the $\log(1+y)$ rate per 100,000 inhabitants to stabilise variance across population sizes.
- **Parallel hybrid (Ensemble)**: an XGBoost regressor [7] on calendar features and lagged and rolling-window statistics (lags $\{1, 2, 4, 8, 13, 26, 52\}$ weeks, spanning short-term momentum, quarterly structure and the annual cycle), blended with Prophet in count space as $\hat{y}_t = \alpha\hat{y}_t^P + (1-\alpha)\hat{y}_t^{XGB}$ with α fitted on out-of-fold predictions, so that Prophet's smooth trend and seasonality complement XGBoost's non-linear lag interactions.
- **Stacked hybrid (Stacking)**: Prophet, ETS and LightGBM [18] experts—two structural and statistical views of trend and seasonality plus one non-linear learner—combined by a non-negative ElasticNet meta-learner over out-of-fold predictions; the non-negativity keeps the learned weights interpretable as positive contributions and stable on short series.

Low-incidence regional fallback. Series with structurally low historical incidence—cumulative count below 5 cases in the 52 weeks immediately preceding the forecast origin—are reassigned to the model selected for their socio-urban region (Fig. 3). On such near-zero series the percentage metrics are dominated by noise, so a regional model estimated from a denser signal is a safer default. The rule depends only on past observations and is therefore free of look-ahead bias. On the depression panel studied here no series falls below the threshold, so the fallback is not activated; it is retained as a safeguard for sparser conditions in the SINAVE catalogue.

3.3 Evaluation Protocol

Cross-validation. Every series–stratum combination is evaluated by rolling-origin cross-validation with five folds, a 52-week horizon and a 13-week (one-quarter) stride; the training window expands at each fold. Metrics are accumulated across folds and weeks, and we report the median across folds per {series, stratum, model} triple.

Metrics and their rationale. Let y_t be the observed value and $\hat{y}_t$ the point forecast at step t over a horizon H. We use four complementary metrics—the root mean squared error (RMSE), the mean absolute error (MAE), the symmetric mean absolute percentage error (SMAPE) and the mean absolute scaled error (MASE)—defined in Eqs. (1)–(2), so that no single artefact dominates the selection.

$$\text{RMSE} = \sqrt{\tfrac{1}{H}\textstyle\sum_{t=1}^{H}(y_t-\hat{y}_t)^2}, \qquad \text{MAE} = \tfrac{1}{H}\textstyle\sum_{t=1}^{H}|y_t-\hat{y}_t|, \tag{1}$$

$$\text{SMAPE} = \frac{100}{H}\sum_{t=1}^{H}\frac{|y_t-\hat{y}_t|}{(|y_t|+|\hat{y}_t|)/2}, \qquad \text{MASE} = \frac{\frac{1}{H}\sum_{t=1}^{H}|y_t-\hat{y}_t|}{\frac{1}{n-52}\sum_{t=53}^{n}|y_t-y_{t-52}|}. \tag{2}$$

SMAPE is scale-normalised and symmetric, hence comparable across entities of very different volume, and is our primary criterion; MASE [15] normalises by a 52-week seasonal-naïve benchmark, so a value below 1 beats seasonal persistence regardless of scale; RMSE penalises large errors and MAE is an interpretable case-count error. Excluding the W53 weeks of the two 53-week years shifts every median MASE by less than 0.005 and leaves the ranking unchanged.

Leakage prevention. Temporal leakage is controlled throughout: the rolling-origin splits are strictly expanding; lag and rolling features never reference an observation later than the forecast origin (multi-step rollout fills $h{>}1$ lags *recursively* with predictions, never future values); demographic covariates are pre-origin annual estimates; and the low-incidence fallback depends only on past counts. The 2026 bulletins of Sect. 5 were used *exclusively* for evaluation: no 2026 observation trained, tuned or reselected any model, and all hyperparameters, architectures and the selection rule were fixed on data through 2025-W53. We call the evaluation *out-of-sample* rather than *blind*, the guarantee being the temporal one above. As a leakage check, an otherwise-identical global versus local (single-series) DeepAR pair on eight states improved the median SMAPE by only $\sim$5% (24.9 vs. 26.3): a modest, legitimate cross-learning gain, not an information leak.[6]

Baselines. Beyond the four candidate models of Sect. 3.2, the 52-week seasonal-naïve benchmark is the denominator of MASE and is reported explicitly (Table 1), together with two baselines re-implemented under the identical CV: a tuned Auto-ARIMA with Fourier annual terms and a PatchTST transformer.

Model selection strategy. For a series–stratum combination, let $\mathcal{E}$ be the candidate set and, for each model $e \in \mathcal{E}$, let s_e, a_e and r_e denote its median cross-validated SMAPE, MASE and RMSE. Write $s_{(1)} \le s_{(2)}$ for the two smallest SMAPE values and $\mathcal{B} = \{e \in \mathcal{E} : s_e \le 1.05\, s_{(1)}\}$ for the models within 5% of the best. The selected model $e^\star$ is determined deterministically:

[6] This controlled pair uses a lightweight identical configuration to isolate the global-versus-local effect; its absolute errors exceed those of the production-configured model in Table 1.

1. **Primary (SMAPE).** If $s_{(2)} > 1.05\, s_{(1)}$ (the best model beats the second by more than a 5% relative margin), then $e^\star = \arg\min_{e \in \mathcal{E}} s_e$.
2. **Tiebreaker (MASE).** Otherwise $e^\star = \arg\min_{e \in \mathcal{B}} a_e$.
3. **Tiebreaker (RMSE).** If the MASE values in $\mathcal{B}$ are tied, $e^\star = \arg\min_{e \in \mathcal{B}} r_e$.

The low-incidence regional fallback of Sect. 3.2 is then applied. The resulting choice and its metrics are persisted alongside each forecast, so every locked prediction is justified by an auditable record rather than an architectural prior. In compact form:

```
sort the four models by cross-validated sMAPE; e1 = best, e2 =
second.
if relative gap(e1, e2) exceeds 5%: select e1.
else: among models within 5% of e1's sMAPE, select the lowest MASE
 (break remaining ties by RMSE).
if cumulative count over the last 52 weeks is below 5: use the
 regional model.
```

Search budget. The per-model search effort is calibrated to its dominant source of sensitivity. Prophet is exposed via a 12-point grid over the changepoint and seasonality priors, the parameters most sensitive to scale across the 111 panels; DeepAR is fixed at the configuration selected on the validation portion of the first cross-validation fold, with context length and dropout the only series-aware choices; the gradient-boosting models use the library defaults with the blend weight α fitted on out-of-fold residuals. We do not equalise the number of configurations across models, because the marginal value of additional DeepAR tuning at panel scale is empirically small relative to its compute cost under the production cross-validation.

Code availability. The model-selection, evaluation and analysis code supporting this study is publicly available at https://github.com/IntegradorIMSS2026Team01/per-series-model-selection2026.

4 Results

4.1 Aggregate Performance

Table 1 reports median per-model accuracy together with two coverage indicators over the 111 combinations: the share in which a model beats the seasonal-naïve baseline (MASE < 1), and the share in which the selection strategy locks it as the forecast. These cross-validation figures drive *selection*; their out-of-sample counterpart, examined below and in Sect. 5, is substantially weaker, so they should be read as a selection signal and not as expected live accuracy.

DeepAR achieves the lowest median value on all four metrics by a wide margin and beats the seasonal-naïve baseline on 99.1% of combinations. Under the selection strategy it is locked for 107 of 111 combinations (96.4%), and for every one of the $32 \times 3 = 96$ federal-entity combinations; the four exceptions

Table 1. Aggregate *cross-validation* performance over the 111 series–stratum combinations (in-sample model-selection metrics, not out-of-sample accuracy; see the generalisation note in the text). The first four numeric columns are the *median* of each accuracy metric across the 111 combinations (lower is better). *Beats naïve* is the share with MASE < 1; *Selected* is the share locked by the strategy of Sect. 3.3. The two rates do not sum to 100% by row because every model is evaluated on every combination but only one is selected per combination.

	Median accuracy (lower is better)				**Coverage (% of 111)**	
Model	**SMAPE (%)**	**MASE**	**RMSE**	**MAE**	**Beats naïve**	**Selected**
Seasonal naïve (52-wk lag)	29.28	1.000	22.57	17.87	—	—
DeepAR	**5.99**	**0.171**	**4.79**	**2.19**	**99.1**	**96.4**
Prophet	29.71	0.885	15.95	12.63	68.5	0.9
Ensemble	25.07	0.925	16.37	12.97	62.2	1.8
Stacking	27.56	0.940	17.02	13.75	55.9	0.9
Additional baselines, re-implemented under the identical CV[‡]						
Auto-ARIMA + Fourier	29.07	1.438	—	—	27.0	—
PatchTST (transformer)	27.18	1.409	—	—	27.0	—

[‡] The Auto-ARIMA baseline models $\log(1+y)$ with an ARIMA order selected by AIC over a compact grid and Fourier annual terms (K=4), the standard treatment of long weekly seasonality; PatchTST [22] is run with its reference configuration (patch length 16) on the same 104-week context window as DeepAR. Both are deliberately out-of-the-box implementations of widely cited baselines, not exhaustively tuned competitors. We report their scale-free SMAPE and MASE, which cross-check against the production pipeline at the national stratum (SMAPE 8.9% vs. 8.8%; MASE 0.60 vs. 0.59); RMSE/MAE are omitted as their absolute scale is not comparable across the two implementations. Neither participates in the model-selection pool.

are aggregate series (the three national strata, discussed next, and one sparse southern region). Because DeepAR's federal-entity margin far exceeds the 5% selection band, this assignment is insensitive to the exact threshold, which only adjudicates the near-ties among the aggregate series.

Two additional baselines evaluated under the identical protocol—a tuned Auto-ARIMA with Fourier seasonality and a PatchTST transformer—land with the statistical models and well above DeepAR's cross-validation value, confirming the gap is not an artefact of a weak classical baseline. A grid over the transformer's patch length and context window does not narrow this gap: across five configurations its general-stratum cross-validated SMAPE stays within 22–23% (best 22.3%), several times the recurrent model's median, so the gap does not stem from an untuned baseline.

Why the recurrent model dominates, and where it does not. At the federal-entity level, where each series is fitted with its own model, two factors favour DeepAR: the flexible autoregressive recurrent structure, with a long context window and a Student-t likelihood, captures each state's seasonality and post-pandemic level

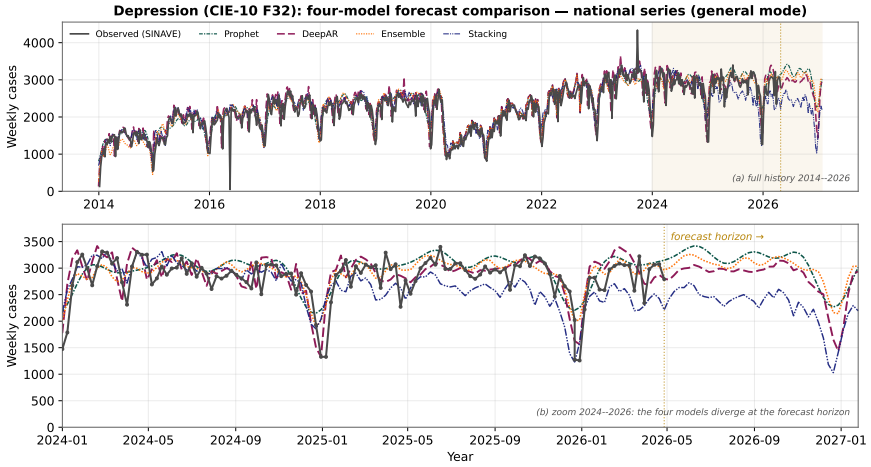

Fig. 4. Four-model forecast comparison on the national series (general stratum). (a) Full 2014–2026 history: the four models track the observed series closely (the gold band marks the window magnified in (b)). (b) Zoom on 2024–2026: past the dotted forecast-horizon line the models diverge, with DeepAR following the magnitude and phase of the seasonal peaks, Prophet under-shooting the post-2022 level and Stacking lagging in phase. At the national level the DeepAR forecast is the joint model summed over states; aggregation smooths state-level variation, so the selection rule retains a simpler model for this stratum.

shift and converges to a strong in-sample fit on this dense, clean panel; and the probabilistic likelihood regularises training on asymmetric counts. We caution that this is an in-sample cross-validation fit that does not persist out of sample (Sect. 4). At the national-aggregate level the advantage is neutralised: there the DeepAR forecast is the joint model summed over states, and aggregation smooths the idiosyncratic state variation, so simpler models fit the smooth trend at least as well. The strategy therefore locks Ensemble for the national general and women strata and Prophet for men, with cross-validated SMAPE of 8.75%, 7.11% and 12.64% respectively (MASE 0.59, 0.46, 0.76); we retain these data-driven choices rather than override them. That the rule lands on different models across the three national strata—a tree-based hybrid for the general and female series and a structural model for the smaller male series—is the per-series mechanism working as intended: each stratum is matched to the inductive bias its own cross-validated evidence supports rather than inheriting the federal-entity winner, so these aggregate-level choices reflect a different data regime, not a contradiction of the federal-entity result. On the national series itself, DeepAR follows the magnitude and phase of the seasonal peaks while Prophet under-shoots the post-2022 level shift and Stacking lags in phase (Fig. 4).

Statistical significance. A one-sided Wilcoxon signed-rank test on the 111 paired per-series SMAPE values confirms that DeepAR lies below each of the other three

candidates in cross-validation ($p < 0.001$ against Prophet, Ensemble and Stacking; n=111), so the cross-validation ranking is not an artefact of the median statistic.

Generalisation to live 2026 data. Table 1 reports in-sample cross-validation metrics, to be read as the basis for *selection*, not as out-of-sample accuracy. Re-scored against the 2026 bulletins, the per-series median SMAPE rises for *every* model and the cross-validation gap closes: the four become comparable (per-series median in the 45–48% range), reflecting both genuine one-year-ahead extrapolation difficulty and the instability of SMAPE on low-count series. The framework is designed for this: per-series assignments are revisable, and re-selecting on the accumulating 2026 signal reassigned 72 of the 111 series; the out-of-sample evaluation resolves this by model and stratum (Sect. 5). Where counts are large and SMAPE stable (the national strata) the locked forecast generalises well, as the prospective evaluation quantifies (6.63% over W02–W18); we treat that, not the cross-validation medians, as the primary out-of-sample evidence of this study.

4.2 State-Level Behaviour

Across the 32 federal entities the selected (DeepAR) model ranges from a best case near 1% SMAPE (Michoacán 1.03, Morelos 1.40, Guanajuato 1.94, Baja California 2.09, Coahuila 2.15) to a worst case in the low double digits (Nuevo León 7.91, Querétaro 9.75, Tlaxcala 11.47, Durango 12.91, Puebla 16.28), with MASE below 0.5 throughout, i.e. a uniform improvement over seasonal persistence. In the worst case (Puebla, MASE = 0.42) the abrupt 2023 level shift inflates the residual variance, but the seasonal phase is still preserved. The per-state out-of-sample errors for every entity, on which the production rule re-selects, are reported in Appendix B.

4.3 Demographic Forecasts

Forecasting the three strata separately surfaces gender-specific saturation windows. The selected national female model (Ensemble, SMAPE = 7.11%) projects 116,047 cases for 2026, against 43,500 male cases and 155,966 in the general stratum; the annual peak falls in February–May for both sexes. Because the strata are forecast independently rather than reconciled, the women-plus-men sum exceeds the general forecast by about 2.3%; we treat this small incoherence explicitly and return to reconciliation in Sect. 6.

5 Out-of-Sample Evaluation (2026)

We assess the locked national-general forecast (Ensemble) against every 2026 SINAVE bulletin published since the training cut-off, covering epidemiological weeks W01–W18 (Table 2, Fig. 5). No 2026 observation entered training, tuning or selection (Sect. 3.3).

Table 2. Out-of-sample evaluation: SINAVE-published weekly incidence (national, general stratum) versus the locked Ensemble forecast for 2026 weeks W01–W18. Deviation is $100(\hat{y} - y)/y$. W01 is reported separately, as the bulletin's first week is structurally incomplete (holiday-season carry-over from the prior W52); aggregates are over the complete weeks W02–W18. The 'Recon.' column is the minimum-trace reconciled national-general forecast (Sect. 6).

Week	Obs.	Pred.	Recon.	Dev. (%)	Week	Obs.	Pred.	Recon.	Dev. (%)
W01*	1,259	2,022	—	+60.6	W10	3,057	3,132	3,190	+2.5
W02	2,220	2,511	2,487	+13.1	W11	3,052	3,004	3,061	−1.6
W03	2,819	2,869	2,835	+1.8	W12	2,574	2,891	2,953	+12.3
W04	2,820	3,044	3,030	+7.9	W13	3,223	2,897	2,955	−10.1
W05	2,595	2,973	3,012	+14.6	W14	2,341	2,961	3,030	+26.5
W06	2,585	2,886	2,961	+11.6	W15	2,981	3,093	3,124	+3.8
W07	2,975	2,906	2,997	−2.3	W16	3,065	3,048	3,112	−0.6
W08	**3,041**	**3,046**	**3,131**	**+0.2**	W17	3,075	3,032	3,080	−1.4
W09	3,086	3,167	3,228	+2.6	W18	2,791	2,964	3,033	+6.2

Cumulative W02–W18: observed 48,300, predicted 50,424 (reconciled 51,219), deviation +**4.40**% (reconciled +6.0%) (∗W01 excluded from aggregates).

Aggregate behaviour. Over the 17 complete weeks W02–W18 the locked forecast attains a SMAPE of **6.63%** (mean MAE = 184 cases per week) and a cumulative deviation of +4.40%, inside the ±5% planning tolerance. The relevant comparator is the CV-period SMAPE on the same national-general stratum, 8.75%: the out-of-sample error is below the in-sample CV error. The weekly errors have a median absolute percentage error of 3.7%, below the 6.6% CV median, so the forecast generalises without visible drift. The ranking is robust to the error metric: under a count-appropriate mean Poisson deviance the locked Ensemble forecast is again the best of the four models on this window (21.7, against 31.9 for Prophet, 32.0 for DeepAR and 87.5 for Stacking). Consistent with the deployment setting, an empirical 80% prediction interval (10th/90th percentiles of post-2021 residuals) covers 15 of 17 validation weeks (88%; 76% on a full-history calibration), bracketing the nominal level rather than the over-confidence of a single model's native intervals.

Pairwise comparison. A Diebold-Mariano test (Harvey-Leybourne-Newbold small-sample correction) on the W02–W18 squared-error losses does not reject equal accuracy between the locked Ensemble and DeepAR ($p = 0.15$) or Prophet ($p = 0.10$), and finds it more accurate than Stacking ($p = 0.03$); with only seventeen weeks the test has limited power, so a non-rejection means insufficient evidence to separate the leading models, not equivalence.

Largest single-week deviation. W14 is the largest single-week deviation (+26.5%), well beyond the upper quartile of the cross-validation error band. The value remained at 2,341 cases through the manuscript cut-off and is therefore not a

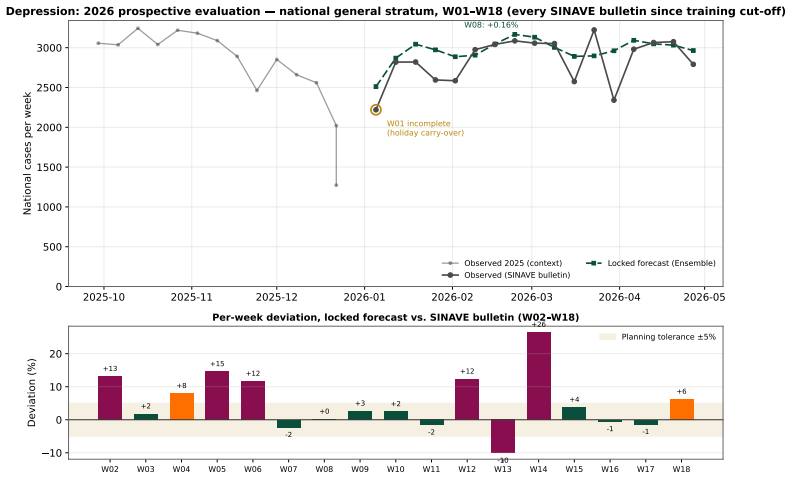

Fig. 5. 2026 out-of-sample evaluation, national general stratum. (a) Late-2025 context and W01–W18 observed (grey) versus the locked Ensemble forecast (green); W01 is flagged as incomplete and the W08 planning milestone is annotated. (b) Per-week deviation against the ±5% planning tolerance band (shaded); bars are coloured by absolute deviation: green within ±5%, orange between 5 and 10%, and wine above 10%.

reporting-lag artefact; the four subsequent weeks W15–W18 return to a SMAPE of 2.9%, so the miss does not propagate into the trajectory. We treat W14 as one observation in the tail of the in-sample error distribution rather than as evidence of model drift, and note that the most recent one or two weeks of any bulletin should still be read with caution because of the documented SINAVE reporting lag. Week W08, the onset of the February–May saturation window, deviates by only +0.2%, though the 17-week aggregate is the more informative summary.

Per-series behaviour. Cross-validation accuracy does not persist per series out of sample (Sect. 4); Fig. 6 resolves this by model and demographic stratum. The cross-validated dominance of the recurrent model (dashed line) does not survive: out of sample no model is uniformly best, and the error grows toward the sparse male stratum, where low weekly counts also inflate SMAPE on small denominators—on the scale-free MASE the same male series stay near 0.45–0.50 for all four models, well below 1, about twice as accurate as 52-week seasonal persistence. Far from undermining per-series selection, this is the regime that makes an *auditable, revisable* rule valuable, and we test that value directly. In a held-out reselection experiment—re-selecting each series' model on the early-2026 weeks (W02–W11) and scoring it on the subsequent, unused weeks (W12–W18)—the revised model lowers held-out SMAPE in 69% of the reassigned series (robust at 61–78% across split points), cutting their median held-out SMAPE from 32.0% to 26.4%. Revising the per-series choice on accumulating data therefore

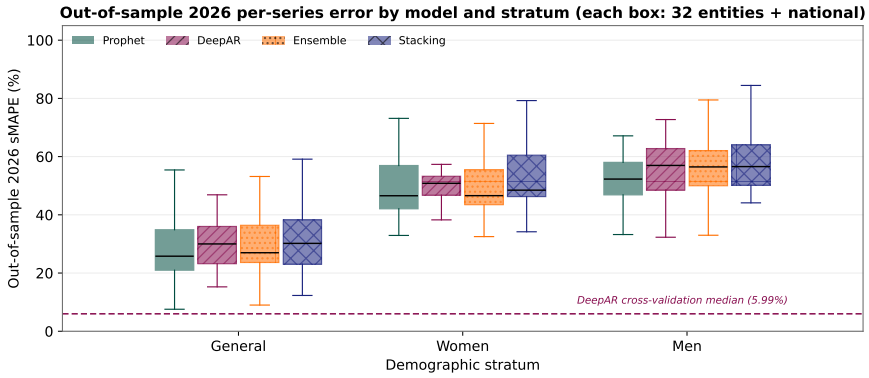

Fig. 6. Out-of-sample 2026 per-series symmetric error by model and demographic stratum; each box summarises the 32 federal entities and the national series over the 2026 weeks available at the production cut-off. In cross-validation the recurrent model is selected for most series at a 5.99% median (dashed line); out of sample that advantage disappears and the four models converge near 45–48%, worst on the sparse male stratum. The auditable rule treats these scores as a revisable signal and reassigns 72 of the 111 series.

measurably improves live accuracy: the rule is a working mechanism, not merely an audit log.

6 Discussion

Combining forecasters with complementary inductive biases and selecting one per series by a transparent, revisable rule is a principled way to forecast a heterogeneous panel, preserving the simpler model where an aggregate series favours it and re-selecting as live data accumulate. The auditable log makes each locked forecast defensible by a metric rather than an architectural preference; we stress that the cross-validation ranking favouring the recurrent model is a selection signal, not an out-of-sample guarantee, since per-series accuracy degrades on live 2026 data (Sect. 4).

Public-health relevance. The 74%/26% reported female–male split, broadly echoed by the 2026 forecast (about 73%/27%), has planning implications: a grid sized at the national mean may underserve women in absolute terms, while male depression remains plausibly under-identified. A sex-disaggregated 52-week forecast could support both proactive capacity allocation for the February–May saturation window and male-targeted help-seeking campaigns in quieter months. We emphasise that the data are *reported incidence*, not prevalence; the female–male gap is a hypothesis requiring dedicated epidemiological validation before use as policy input. This paper evaluates the framework; it does not describe a deployed operational system.

Hierarchical reconciliation. The three demographic strata are forecast independently, leaving the ∼2.3% general-versus-(women+men) incoherence noted in Sect. 4.3. Applying a minimum-trace reconciliation [27] to the locked national forecasts, weighted by the per-stratum cross-validation error variances, removes the incoherence by construction; the reconciled national-general series (Table 2, 'Recon.' column) shifts the operational forecast by only 1.9% on average and, since it is already well calibrated, slightly raises the cumulative deviation (+6.0% vs. +4.40%). The independence approximation is thus a small source of error, and we report the unreconciled forecasts for transparency.

Limitations. Five limitations bound the conclusions. (i) Only one condition (depression) is evaluated; transfer to conditions with different temporal behaviour remains to be shown. (ii) The per-series cross-validation metrics of Table 1 are optimistic relative to live performance: re-scored on the 2026 bulletins the per-series errors rise and the cross-validation ranking compresses (Sect. 4). The strong cross-validation result is best read as the basis for *selection*, and the out-of-sample claim of this paper rests on the national-aggregate prospective evaluation, not on the cross-validation medians. (iii) The headline prospective evaluation is at the national-aggregate stratum over an early-2026 window (W01–W18); a per-state out-of-sample table over the same window is given in Appendix B, but a per-state *full-year* evaluation is still pending. (iv) The most recent bulletins are subject to reporting-lag revision, so the latest one or two weeks are provisional. (v) The framework has not been externally validated outside SINAVE/Mexico.

Future work. We plan to extend the framework to further CIE-10 codes in the SINAVE catalogue, to report a full-year per-state prospective evaluation, to add tuned SARIMA and N-BEATS/N-HiTS baselines, and to pursue external validation.

7 Conclusions

We presented an empirical study of multi-model, per-series forecast selection for weekly depression incidence at federal-entity granularity in Mexico, comparing a structural model, a global probabilistic recurrent network and two tree-based hybrids under a deterministic, auditable and revisable selection rule. Evaluated on more than a decade of SINAVE data across a nation-wide panel of series and demographic strata, cross-validation selects the recurrent model (DeepAR) for the large majority of federal-entity series, while the rule retains simpler models where aggregation favours them. We are explicit that this ranking is an in-sample selection signal: on live 2026 data the per-series advantage narrows and the framework re-selects, so our out-of-sample evidence is the prospective national evaluation, not the cross-validation figures. We are not aware of prior work that applies a multi-model framework with auditable, revisable per-series model selection to psychiatric surveillance at federal-entity granularity in a Latin American country. The framework is a research prototype; future work will broaden the conditions, baselines and prospective horizon evaluated.

Ethics and Data Availability

The SINAVE bulletin reports aggregate weekly case counts at the federal-entity level, stratified by sex; all analyses were performed on these aggregate cells and no cell corresponds to an individual patient, so no informed consent or review-board approval was required. The institutional ethics committee of Tecnológico de Monterrey (ITESM-HGO) confirmed that no formal review was required for the use of aggregate, public SINAVE data. The data are published as open data by the Dirección General de Epidemiología, Secretaría de Salud, and were retrieved from the public SINAVE portal on 2026-04-15 [11]; redistribution is governed by the open-data terms of the bulletin.

Acknowledgments. The authors gratefully acknowledge subject-matter guidance from Dra. Ruth Manuela Pérez-Hernández (IMSS) and Dra. Grettel Barceló-Alonso (ITESM-HGO) on the model-selection methodology, and the contributions of Dra. Lina Díaz-Castro (Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz) and Dra. Silvia Magali Cuadra-Hernández (Instituto Nacional de Salud Pública) on the epidemiological interpretation of the data. The authors thank the M.Sc. in Applied Artificial Intelligence (MNA) programme of Tecnológico de Monterrey for institutional support.

Disclosure of Interests. The authors have no competing interests relevant to the content of this article. The corporate affiliations in the header indicate current employment only; this work was developed independently of those employers under the MNA programme, with no funding, computing resources or proprietary information from either company.

A Implementation and Reproducibility Details

The four models are exposed behind a single, common interface and configured from hierarchical YAML files under OmegaConf. The framework is implemented in Python 3.12 on GluonTS 0.15 (PyTorch 2.5), Prophet 1.1, XGBoost 2.0, LightGBM 4.3, scikit-learn 1.4 and pandas 2.2. The recurrent model (DeepAR) is trained on a single NVIDIA T4 GPU (an `ml.g4dn.xlarge` cloud instance), while the remaining models run on CPU. Run-to-run reproducibility is enforced with a fixed multi-source seed: `torch`, `numpy` and `random` are set to 42, `torch.backends.cudnn.deterministic` is enabled, and the per-library `feature_fraction_seed` and `bagging_seed` of the tree-based models are fixed. Per-model hyperparameters are summarised in Table 3, and the configuration files and the model-selection pseudocode are released with the source code.

B Per-State Out-of-Sample Evaluation (2026)

Table 4 reports the per-entity out-of-sample 2026 SMAPE of the four models (general stratum); its wide spread—genuine state heterogeneity compounded by SMAPE inflation on sparse counts—is what the per-series, revisable rule is built to absorb, and is why no single model dominates the panel, rather than a shortcoming of any one forecaster.

Table 3. Principal hyperparameters per model, fixed before the 2026 evaluation window.

Model	Principal hyperparameters
DeepAR	2 layers × 80 cells; context 104 wk; horizon 52; dropout 0.15; Student-t output; Adam η=5×10^{-4}, λ_2=10^{-6}; batch 32; $\leq$300 epochs; early stop (patience 15).
Prophet	rate per 100k, $\log(1+y)$; changepoint prior $\in \{0.01, 0.03, 0.05\}$; seasonality prior $\in \{0.025, 0.05, 0.1, 0.5\}$ (12-point grid under temporal CV).
Ensemble	XGBoost on calendar/lags $\{1, 2, 4, 8, 13, 26, 52\}$/rolling stats; blend weight learned on out-of-fold predictions.
Stacking	Prophet + ETS + LightGBM experts; non-negative ElasticNet (α=1.0, `l1_ratio`=0.5).

References

1. Albert, P.R.: Why is depression more prevalent in women? J. Psychiatry Neurosci. **40**(4), 219–221 (2015). https://doi.org/10.1503/jpn.150205
2. Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., et al.: GluonTS: probabilistic and neural time series modeling in Python. J. Mach. Learn. Res. **21**(116), 1–6 (2020). http://jmlr.org/papers/v21/19-820.html
3. Ansari, A.F., Stella, L., Türkmen, C., Zhang, X., Mercado, P., Shen, H., et al.: Chronos: learning the language of time series. Trans. Mach. Learn. Res. (2024). https://doi.org/10.48550/arXiv.2403.07815
4. Benidis, K., Rangapuram, S.S., Flunkert, V., Wang, Y., Maddix, D., Türkmen, C., et al.: Deep learning for time series forecasting: tutorial and literature survey. ACM Comput. Surv. **55**(6), 1–36 (2022). https://doi.org/10.1145/3533382
5. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th edn. John Wiley & Sons, Hoboken (2015)
6. Challu, C., Olivares, K.G., Oreshkin, B.N., Garza Ramirez, F., Mergenthaler Canseco, M., Dubrawski, A.: N-HiTS: neural hierarchical interpolation for time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 6, pp. 6989–6997 (2023). https://doi.org/10.1609/aaai.v37i6.25854
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794. ACM, San Francisco (2016). https://doi.org/10.1145/2939672.2939785
8. Chimmula, V.K.R., Zhang, L.: Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals **135**, 109864 (2020). https://doi.org/10.1016/j.chaos.2020.109864
9. Cramer, E.Y., Ray, E.L., Lopez, V.K., Bracher, J., Brennen, A., Castro Rivadeneira, A.J., et al.: Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc. Natl. Acad. Sci. USA **119**(15), e2113561119 (2022). https://doi.org/10.1073/pnas.2113561119

Table 4. Per-state out-of-sample 2026 SMAPE (%), general stratum, weeks W02–W18 (Dpr DeepAR, Prp Prophet, Ens Ensemble, Stk Stacking); **bold** marks each series' lowest error. Cross-validation selects DeepAR for all 32 entities (Ensemble nationally), yet out of sample it is lowest in only 10 of the 33 series. Medians: DeepAR 30.0, Prophet 25.8, Ensemble 27.0, Stacking 30.2. The bold model is the rule's pick by construction; the held-out test of Sect. 5 (scored on *unused* weeks) shows the reassignment generalises.

Entidad	Dpr	Prp	Ens	Stk	Entidad	Dpr	Prp	Ens	Stk
Aguascalientes	26	**20**	25	21	Nayarit	**30**	34	38	38
Baja California	20	**8**	9	12	Nuevo León	31	38	**30**	37
Baja Calif. Sur	39	49	40	**37**	Oaxaca	**22**	26	31	30
Campeche	37	26	26	**25**	Puebla	46	**37**	53	59
Chiapas	35	35	**33**	37	Querétaro	43	**32**	33	35
Chihuahua	**15**	17	16	**15**	Quintana Roo	**28**	35	**28**	31
Ciudad de México	26	**20**	24	39	San Luis Potosí	34	**32**	36	49
Coahuila	36	**23**	24	30	Sinaloa	31	**25**	35	46
Colima	**30**	47	32	72	Sonora	23	21	22	**20**
Durango	34	31	**27**	29	Tabasco	**47**	64	52	61
Guanajuato	**21**	39	42	36	Tamaulipas	27	**23**	25	30
Guerrero	40	25	25	**23**	Tlaxcala	**44**	55	53	72
Hidalgo	23	**20**	22	23	Veracruz	26	28	26	**22**
Jalisco	21	**14**	15	16	Yucatán	31	25	37	**24**
Michoacán	**40**	63	68	92	Zacatecas	**16**	23	21	23
Morelos	23	**21**	24	24	**Nacional**	16	**13**	**13**	18
México	24	23	**19**	33					

10. Dash, S., Giri, S.K., Mallik, S., Pani, S.K., Shah, M.A., Qin, H.: Predictive healthcare modeling for early pandemic assessment leveraging deep auto regressor neural prophet. Sci. Rep. **14**, 5287 (2024). https://doi.org/10.1038/s41598-024-55973-y
11. Dirección General de Epidemiología (DGE), Secretaría de Salud, México: Sistema Nacional de Vigilancia Epidemiológica (SINAVE) — Boletín Epidemiológico. https://www.gob.mx/salud/acciones-y-programas/direccion-general-de-epidemiologia-boletin-epidemiologico, last accessed 2026-04-15
12. García-Pacheco, J.Á., Torres-Ortega, M.L., Borges, G.: The burden of mental disorders in Mexico, 1990–2019: mental and neurological disorders, substance use, suicides, and related somatic disorders. Span. J. Psychiatry Ment. Health **17**(2), 64–70 (2024). https://doi.org/10.1016/j.rpsm.2021.06.005
13. Godahewa, R., Bergmeir, C., Webb, G.I., Montero-Manso, P.: An accurate and fully-automated ensemble model for weekly time series forecasting. Int. J. Forecast. **39**(2), 641–658 (2023). https://doi.org/10.1016/j.ijforecast.2022.01.008
14. Hewamalage, H., Bergmeir, C., Bandara, K.: Recurrent neural networks for time series forecasting: current status and future directions. Int. J. Forecast. **37**(1), 388–427 (2021). https://doi.org/10.1016/j.ijforecast.2020.06.008
15. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. Int. J. Forecast. **22**(4), 679–688 (2006). https://doi.org/10.1016/j.ijforecast.2006.

03.001
16. Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D.: Forecasting with Exponential Smoothing: The State Space Approach. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-71918-2
17. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice, 2nd edn. OTexts, Melbourne (2018). https://otexts.com/fpp2/
18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146–3154 (2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
19. Lim, B., Zohren, S.: Time-series forecasting with deep learning: a survey. Philos. Trans. R. Soc. A **379**(2194), 20200209 (2021). https://doi.org/10.1098/rsta.2020.0209
20. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 Competition: 100,000 time series and 61 forecasting methods. Int. J. Forecast. **36**(1), 54–74 (2020). https://doi.org/10.1016/j.ijforecast.2019.04.014
21. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M5 accuracy competition: results, findings, and conclusions. Int. J. Forecast. **38**(4), 1346–1364 (2022). https://doi.org/10.1016/j.ijforecast.2021.11.013
22. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR) (2023). https://doi.org/10.48550/arXiv.2211.14730
23. Reich, N.G., Brooks, L.C., Fox, S.J., Kandula, S., McGowan, C.J., Moore, E., et al.: A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc. Natl. Acad. Sci. USA **116**(8), 3146–3154 (2019). https://doi.org/10.1073/pnas.1812594116
24. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecast. **36**(3), 1181–1191 (2020). https://doi.org/10.1016/j.ijforecast.2019.07.001
25. Santomauro, D.F., Mantilla Herrera, A.M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D.M., et al.: Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet **398**(10312), 1700–1712 (2021). https://doi.org/10.1016/S0140-6736(21)02143-7
26. Taylor, S.J., Letham, B.: Forecasting at scale. Am. Stat. **72**(1), 37–45 (2018). https://doi.org/10.1080/00031305.2017.1380080
27. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J.: Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J. Am. Stat. Assoc. **114**(526), 804–819 (2019). https://doi.org/10.1080/01621459.2018.1448825
28. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992). https://doi.org/10.1016/S0893-6080(05)80023-1