Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Catalogue attributes versus early accelerograms for recovering retrospective seismic-alert labels in Mexico

María del Carmen Ríos-Perea
Universidad Autónoma Metropolitana
Benjamin Moreno-Montiel
bmm@xanum.uam.mx
Universidad Autónoma Metropolitana
René MacKinney-Romero
Universidad Autónoma Metropolitana

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11736605/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Catalogue attributes versus early accelerograms for recovering retrospective seismic-alert labels in Mexico

María del Carmen Ríos Perea[1], Benjamin Moreno-Montiel[2*], René MacKinney-Romero[3]

[1]Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Unidad Iztapalapa, Av. San Rafael Atlixco 186, Leyes de Reforma 1ra Secc., Ciudad de México, 09340, Mexico.

*Corresponding author(s). E-mail(s): bmm@xanum.uam.mx;
Contributing authors: maryrpcn09@gmail.com; rene@xanum.uam.mx;

Abstract

The Mexican Seismic Alert System (SASMEX) issues public warnings from magnitude estimates computed in the first seconds of strong motion, but the published evaluations of its algorithms used fewer than 100 earthquakes each. We audited 7,562 accelerograms recorded in Mexico between 1964 and 2020 and labelled each record with a retrospective rule derived from the SASMEX dissemination criteria, which combines catalogue magnitude with epicentral distance to a reference point in Mexico City. Classifiers supplied with catalogue magnitude and distance mostly reproduce that rule: a decision tree trained on 2000–2020 records matched 99.40% of the 1964–1999 labels by learning its thresholds, whereas naive Bayes and nearest neighbours fell from out-of-fold accuracies above 97% to balanced accuracies of 61–69%. We then computed amplitude and energy features from 1–3 s of acceleration after an automatic trigger, trained on 2,630 records from 1964–1999 and tested on 4,018 from 2000–2020. With vertical features, 19 classifiers reached balanced accuracies of 49.37–52.50%. At 3 s, three-component features raised balanced accuracy from 50.60% to 55.50% for adaptive boosting (paired gain 4.90 percentage points; 95% interval 2.74–7.14), but the model still missed three quarters of the positive labels. Correcting 101 labels affected by extraction errors, using an improved trigger or starting windows at a reference P onset kept balanced accuracy between 45.66% and 58.72%. High scores obtained with catalogue inputs should therefore not be read as evidence of early-warning skill.

Keywords: Earthquake early warning, Strong motion, SASMEX, Temporal validation, Machine learning, Mexico

Article Highlights

- Scores built from magnitude and distance mostly copy the alert rule and overstate early-warning skill.
- Models trained on older Mexican records and using a few seconds at one station missed most later alert records.
- Fixing data and trigger errors barely helped; distant, noisy Mexico City stations may limit what early shaking shows.

1 Introduction

Mexico lies at the junction of the North American, Cocos, Rivera, Pacific and Caribbean plates, on

the Pacific Ring of Fire. Subduction of the Cocos and Rivera plates along the Pacific coast produces large interplate and intraslab earthquakes whose waves reach Mexico City and other cities of central Mexico tens of seconds after rupture. The 1957 Acapulco (M 7.8), 1985 Michoacán (M_w 8.1) and 2017 Puebla–Morelos (M_w 7.1) earthquakes, which caused severe damage in Mexico City, illustrate the risk.

Earthquake early warning (EEW) systems detect an earthquake near its source, estimate its size or the expected shaking from the first seconds of signal, and warn distant sites before strong shaking arrives (Kanamori 2005; Allen and Melgar 2019). Mexico operates one of the oldest public EEW systems, the Mexican Seismic Alert System (SASMEX) (Espinosa-Aranda et al. 1992, 2009). Its field stations estimate magnitude with three algorithms. The $2(t_S - t_P)$ algorithm relates the energy accumulated in the vertical and horizontal components up to twice the S–P interval to body-wave magnitude (Cuéllar et al. 2017); the $(t_S - t_P)$ algorithm relates the accumulated and peak squared vertical acceleration between the P and S arrivals to moment magnitude (Cuéllar et al. 2018a); and the $(t_P + 3\,\mathrm{s})$ algorithm uses the accumulated squared vertical acceleration and its growth rate in the first 3 s after the P arrival, so it does not wait for the S wave (Cuéllar et al. 2018b).

The published evaluations of these algorithms have four limitations that motivated this work (Table 1). They used small sets of earthquakes, between 24 and 89 per algorithm, which is few for a country with frequent moderate and large earthquakes. They report performance mainly as the percentage of correct decisions, without precision, sensitivity, false-alarm rates or receiver operating characteristic (ROC) analysis (Fawcett 2006). They use different magnitude scales (m_b for $2(t_S - t_P)$, M_w for the other two), which hampers comparison. And they do not use the large public archive of Mexican accelerograms in the ASA 2.0 format: the database of the Instituto de Ingeniería of the Universidad Nacional Autónoma de México (RAII-UNAM) alone offers about 18,600 records (Instituto de Ingeniería, UNAM 2023).

Such archives are a natural input for data-driven alert models, and machine learning is increasingly used for EEW and phase picking (Kong et al. 2016; Mousavi et al. 2020; Münchmeyer et al. 2021). Two difficulties are specific to archives that span decades. Seismicity, instrumentation and data quality change with time, so models must be evaluated on periods not used for training (Gama et al. 2014); and the target must be defined so that it is not trivially available from the inputs.

This paper revisits and extends SASTREMORS (supervised learning system with temporal information for an efficient response in seismic monitoring), an earlier, unpublished study by the authors that built a labelled database from the RAII-UNAM archive, tre-data 1.0, and compared classifiers trained on catalogue attributes, including the magnitude and distance used to construct the labels; a decision tree matched 99.40% of the labels of an older test period. Because the labels are a deterministic function of those two attributes, that result measures how well a classifier rediscovers a known rule, not how much information an early-warning system could extract from the signal. Here we (i) describe and audit the archive and the retrospective labels; (ii) report, as a diagnostic, the cross-validation and temporal-transfer results of the catalogue-attribute classifiers; and (iii) evaluate how much of the same label can be recovered from short causal windows of acceleration after an automatic trigger, training on older records and testing on later ones. Section 2 summarises the SASMEX criteria and algorithms, Section 3 the data and labels, Section 4 the two experiments, Section 5 the results and Section 6 their interpretation and limitations.

2 SASMEX decision criteria and magnitude algorithms

SASMEX started operating in 1991 with twelve strong-motion field stations in the Guerrero gap, equipped with triaxial sensors sampled at 50 Hz and an STA/LTA detector (short-term average over long-term average) (Allen 1978; Espinosa-Aranda et al. 2009). At the time of that study, it had 96 sensing stations along the Pacific coast, from Jalisco to Oaxaca, and in central Mexico, and seven alert dissemination centres (Mexico City, Oaxaca, Acapulco, Chilpancingo, Morelia, Puebla and Cuernavaca). An alert is disseminated when two stations agree: to the centres within 250 km if

both estimate a magnitude of 5.0, to those within 350 km if both estimate 5.5, and to all centres if both estimate 6.0 (Espinosa-Aranda et al. 2009).

Each station detects the P and S arrivals, t_P and t_S, from the average squared acceleration of the vertical and horizontal components over 16-sample windows,

$$\mathrm{ASI}_V(i) = \tfrac{1}{16}\sum_{n=i-15}^{i} x_v^2(n), \quad (1)$$

$$\mathrm{ASI}_H(i) = \tfrac{1}{16}\sum_{n=i-15}^{i} \left[x_l^2(n) + x_t^2(n)\right], \quad (2)$$

and from the ratio of the accumulated vertical and horizontal energies, $V(i) = \sum_{n\le i} x_v^2(n)$ and $H(i) = \sum_{n\le i}[x_l^2(n) + x_t^2(n)]$, weighted by coefficients α_k and β_k in $(0, 1)$ that are calibrated for each station (Cuéllar et al. 2017). P-wave energy dominates the vertical component and S-wave energy the horizontal components, so the weighted ratio $\alpha_k V/(\beta_k H)$ exceeds one after the P arrival and falls after the S arrival. Dynamic thresholds computed from the pre-event noise and the early coda complete the detector, and the S wave must be found within 24 s of the P wave.

The $2(t_S - t_P)$ algorithm computes

$$a = \log \sum_{i=t_P}^{2(t_S-t_P)} \left[\mathrm{ASI}_V(i) + \mathrm{ASI}_H(i)\right], \quad (3)$$

$$m = \log \left[\mathrm{ASI}_V + \mathrm{ASI}_H\right]_{i=2(t_S-t_P)}, \quad (4)$$

and classifies the event with three linear rules: $M \ge 5.0$ if $a + m - 7 \ge 0$, $M \ge 5.5$ if $a + 0.98m - 7.18 \ge 0$ and $M \ge 6.0$ if $a + m - 7.6 \ge 0$ (Cuéllar et al. 2017). The $(t_S - t_P)$ algorithm estimates M_w with a piecewise regression on $\mathrm{SA} = \log_{10}\sum_{t_P}^{t_S} a_v^2(i)$ and $\log_{10}\max_{t_P\le i\le t_S} a_v^2(i)$, where a_v is the vertical acceleration (Cuéllar et al. 2018a). The $(t_P + 3\,\mathrm{s})$ algorithm uses the accumulated squared vertical acceleration $a_{v(T)} = \sum_{t_P}^{t_P+3\,\mathrm{s}} a_v^2(i)$ and an angle θ_P that describes how fast that sum grows in the sub-windows 0–0.5, 0.5–1.75 and 1.75–3 s after t_P, again with a piecewise regression to M_w (Cuéllar et al. 2018b); it was designed for events deeper than 40 km.

Table 1 summarises the published evaluations. They were obtained on post facto sets of a few dozen earthquakes, and performance is reported as the fraction of events whose magnitude class or alert decision was correct.

Table 1 Published evaluations of the SASMEX magnitude algorithms (Cuéllar et al. 2017, 2018a,b)

Algorithm	Earthquakes	Correct	Scale
$2(t_S - t_P)$	61[1]	92%	m_b
$(t_S - t_P)$	89[2]	79 (88.8%)	M_w
$(t_P + 3\,\mathrm{s})$	24[3]	19 (79.2%)	M_w

[1]144 accelerograms; the classification error was larger for $M > 5.5$.

[2]Regression fitted on 324 accelerograms of 101 earthquakes from 1985–2017; evaluation events had at least two stations within 70 km; 6 over- and 4 underestimated.

[3]Events of 1995–2017; 4 over- and 1 underestimated.

Table 2 Blocks of the ASA 2.0 format and the attributes extracted for tre-data 1.0

Block	Content (attributes used in bold)
File	Format, file name, reference, creation time
Station	Name, code, soil type, **coordinates**
Instrument	Model, serial number, channels and orientation, sampling rate, **natural frequency** and **damping** per channel
Earthquake	Date, origin time, **magnitude**, **epicentre**, **focal depth**
Accelerogram	Sampling interval, duration, number of samples, **peak acceleration** per channel
Comments	Quality notes
Data	Acceleration series, one column per channel

3 Data and retrospective labels

3.1 Strong-motion archive and ASA 2.0 attributes

The RAII-UNAM network transfers the records of its accelerographs to binary files, checks that each trigger corresponds to an earthquake and that the instrument worked correctly, converts the record to ASCII and catalogues it (Instituto de Ingeniería, UNAM 2023). Records are distributed in the ASA 2.0 format, a self-describing ASCII file whose header blocks describe the station, the instrument, the earthquake and the accelerogram, followed by the acceleration time series of the three components (Table 2).

From each header, the original study extracted twelve attributes: focal depth; a distance D computed from epicentral coordinates with the spherical law of cosines; the natural frequency and the damping of the sensor in each of the three channels; the signed peak acceleration in each channel; and the catalogue magnitude. The archive used here contains 7,562 accelerograms: 3,487 in a nominal 1964–1999 group and 4,075 in a nominal 2000–2020 group (Fig. 1a).

We associated every archive member with a matrix row. For 7,517 rows, focal depth, distance, natural frequencies, peak accelerations and magnitude identified a single file. For the remaining 45 rows, whose matrix attributes were incomplete, we used file order. Recomputing the attributes from the headers showed how the legacy extractor worked. It took M as the first magnitude listed in the header, whatever its scale, which reproduces the stored value in 7,531 of the 7,562 records. It computed D with the spherical law of cosines, an Earth radius of 6,378 km and a fixed reference point (Section 3.2), taking longitude as the negative of the printed value. This reproduces D within 0.5 km in 7,548 of the 7,557 records that report coordinates.

The audit also found anomalous values that the original study did not filter: 41 recent and 10 older records have a stored magnitude of zero, 32 have a natural frequency of zero, 168 recent records report 500 Hz, and 270 recent and 4 older records have $D > 3{,}000$ km (Fig. 1b). All the magnitude and distance anomalies are reading errors of the legacy extractor, not catalogue values (Table 3). We did not verify the instrument values. Correcting the extraction changes 101 labels in six events, all from no alert to alert and all in 2000–2020. Four older headers contain no earthquake data, so their labels remain undetermined. The corrected labels use the first non-zero magnitude in the header. Because some events list different magnitudes at different stations (343 records), we assigned each event a single label, computed from the median M and D of its records. This changes three further labels: two recent records become alerts and one older record becomes a non-alert. The corrected labels give 705 positive records in 1964–1999 and 2,109 in 2000–2020, against 706 and 2,006 historical positives. With the maximum listed magnitude instead of the first, 182 labels would differ. We kept the historical labels in Experiment 1 to remain comparable with the original study, and we report Experiment 2 with both the historical and the corrected labels (Section 5.5).

3.2 Retrospective alert labels

ASA 2.0 files contain no alert decision, so the original study labelled each record with a rule adapted from the SASMEX dissemination criteria. A record is positive (alert) when

$$(M > 5.0 \wedge D \leq 250) \vee (M > 5.5 \wedge D \leq 350) \vee (M > 6.0 \wedge D > 350), \tag{5}$$

and negative otherwise (Fig. 1b). Here M is the magnitude stored in the historical matrix. Headers report more than one magnitude scale, so we refer to it as catalogue magnitude. D, in kilometres, is the great-circle distance from the epicentre to the fixed point 19.25° N, 99.13° W used by the legacy extractor, which lies in Mexico City; the original study described it as the epicentre–station distance, and the audit corrects that description. The label therefore depends only on the event, so all records of an earthquake share it, and it describes whether the event met a Mexico City-oriented version of the dissemination criteria. It is not an operational SASMEX decision: operational magnitude estimates and alert logs were unavailable for this study.

The two periods differ strongly in class balance and in their attributes (Table 4, Fig. 1). The 2000–2020 records are nearly balanced, whereas only about one fifth of the 1964–1999 records are positive, with either version of the labels. Recent records have much larger catalogue magnitudes (median 5.6 against 4.6; 91% against 38% with $M \geq 5$) and twice the share of intermediate-depth events (20% against 9% at 40–60 km). Sensor natural frequencies cluster around 20–40 and 45–60 Hz in nearly the same proportions in both periods; only recent records report 500 Hz.

4 Methods

Figure 2 summarises the two experiments. Experiment 1 reproduces the original SASTREMORS design with catalogue attributes and serves as a diagnostic. Experiment 2 removes the catalogue

Table 3 Reading errors of the legacy extractor found by recomputing M and D from the ASA 2.0 headers, and their effect on the retrospective labels

Error	Records	Effect on stored values	Labels changed
Longitude printed with a minus sign and "W" and read as east longitude	260	$D \approx 15{,}000$ km	73 (4 events)
Empty first magnitude ("Mb=0.0/Ms=0.0/Mc=6.7")	34	$M = 0$	21 (1 event)
Matrix row with zero instrument fields although the header is complete (event of 2015-03-20)	7	$M = 0$, $D = 9{,}091$ km	7 (1 event)
Matrix row with zero instrument fields (event of 2015-04-28)	2	$D = 9{,}189$ km	0
Impossible latitude (98.00° N)	1	$D > 3{,}000$ km	0
One-letter magnitude scale ("M=3.60" read as 0.6)	15	M too small; $M = 0$ in 3	0
Missing "=" after the scale ("Mb4.0")	3	$M = 0$	0
Unlabelled coordinates	1	–	0
Header without earthquake data (1964–1999)	4	$M = 0$, $D > 3{,}000$ km	undetermined

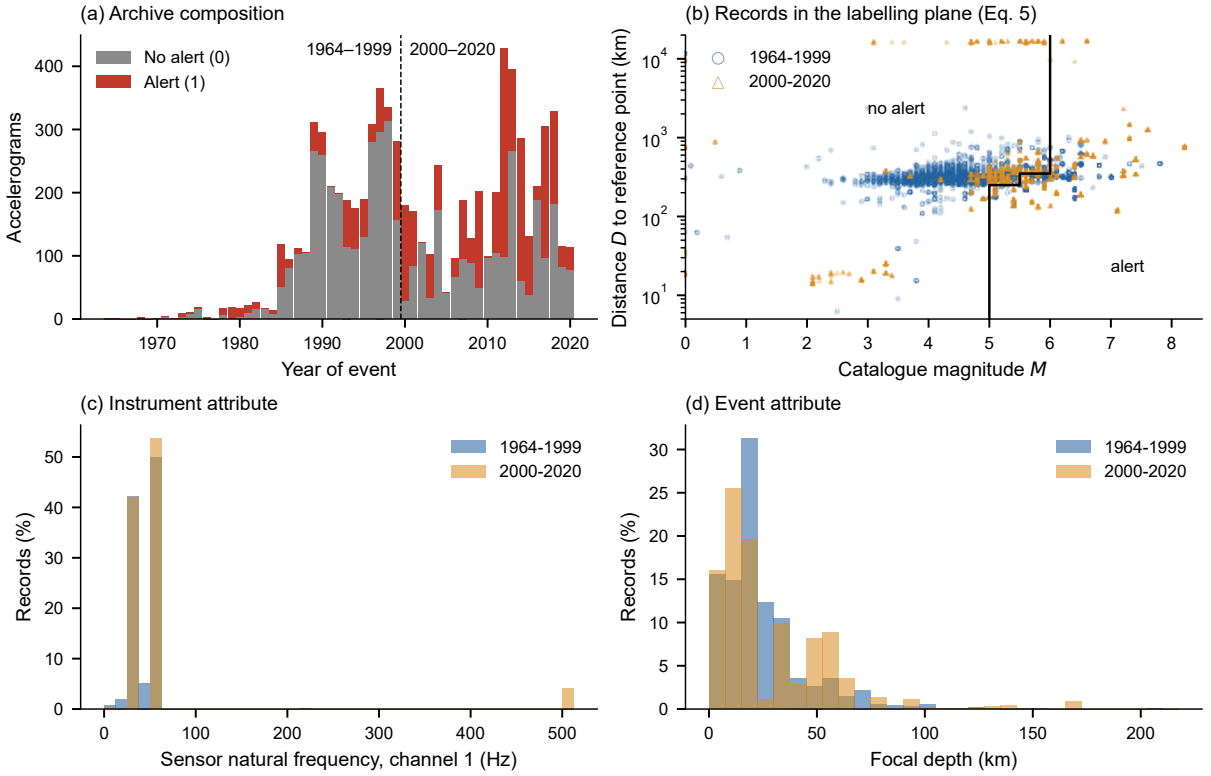

Fig. 1 The tre-data 1.0 archive. (a) Accelerograms per event year and retrospective label; the dashed line separates the two nominal periods. (b) Records in the plane of catalogue magnitude and distance D from the epicentre to the reference point in Mexico City; the black line is the boundary of Eq. 5 (alert to the right and below). Records with $M = 0$ have missing magnitude; records with $D > 3{,}000$ km are discussed in the text. (c) Natural frequency of the sensor in channel 1 and (d) focal depth, as percentages of each period

attributes and uses only early waveform features, with a forward chronological split.

4.1 Experiment 1: catalogue attributes

Experiment 1 uses the twelve attributes of Section 3.1. Following the original study, models were developed on the nearly balanced 2000–2020 matrix and tested on the imbalanced 1964–1999 matrix. A two-dimensional t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton 2008) of the z-scored attributes shows that, in this attribute space, the classes form many small islands rather than two separable groups,

Table 4 Class distribution of the archive with the historical labels stored in the feature matrices and with the corrected labels of Section 3.1. Percentages refer to the labelled records of each period

		Historical		Corrected		
Period	Records	Alert	No alert	Alert	No alert	Undetermined
1964–1999	3,487	706 (20.2%)	2,781	705 (20.2%)	2,778	4
2000–2020	4,075	2,006 (49.2%)	2,069	2,109 (51.8%)	1,966	0

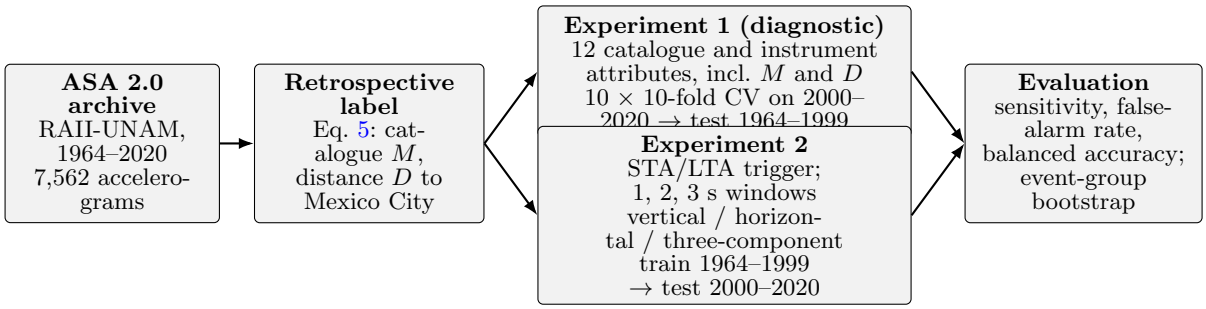

Fig. 2 Workflow of the two experiments. Experiment 1 supplies the classifiers with the attributes used to construct the label; Experiment 2 uses only acceleration recorded after an automatic trigger

and that the islands differ between the periods (Fig. 3).

Five classifiers were implemented from scratch in MATLAB, chosen from the most widely used families (Sarker 2021) and for their short classification time. The decision tree follows C4.5 (Quinlan 1993) in choosing, at each node, the attribute and binary threshold with the largest information gain, the reduction of the class entropy $H(S) = -\sum_c p_c \log_2 p_c$; a minimum leaf size and cost-complexity pruning limit its growth. We call it the legacy tree, because its entropy routine indexes classes from one while the stored labels are 0 and 1, so it is not a validated C4.5 implementation. Decision tables (Kohavi 1995) rank the attributes by their information about the class (entropy minus conditional entropy), discretise each attribute into ten intervals and assign the class of the matching rule; in the archived code the interval edges are recomputed from the batch being classified. The backpropagation neural network (BP-ANN) (Rumelhart et al. 1986) has 13 inputs (the 12 attributes plus a constant bias input), three hidden layers of ten sigmoid units and one output unit, and was trained for a fixed budget of 8,000 epochs; its learning rate, $\lambda = 0.015$, was selected from 500 values drawn from a normal distribution with mean 0.01 and standard deviation 0.005. Naive Bayes (Domingos and Pazzani 1997) estimates conditional probabilities from the frequencies of exact attribute values, and k-nearest neighbours (k-NN) (Cover and Hart 1967) uses the Euclidean distance on z-scored attributes, with k chosen among the odd values from 3 to 21. Six classifiers of the MATLAB Statistics and Machine Learning Toolbox with default settings completed the list of Sarker (2021): linear discriminant analysis (LDA), logistic regression, a support vector machine (SVM) (Cortes and Vapnik 1995), a random forest (Breiman 2001), AdaBoostM1 (Freund and Schapire 1997) and a linear model trained by stochastic gradient descent (SGD).

Each from-scratch classifier was evaluated on the training period with ten repetitions of ten-fold cross-validation; we report the fold accuracies of the best repetition, which are out-of-fold estimates. The model of the best fold was then applied to the 1964–1999 records, which were never used for fitting or tuning. The archived scripts also score that model on the entire training period; those scores include the records used for fitting and are not reported. Finally, as in the original study, the legacy tree was cross-validated on the merged 1964–2020 matrix.

4.2 Experiment 2: automatic trigger and early windows

For Experiment 2 we read the acceleration series of every record, subtracted the first sample of the

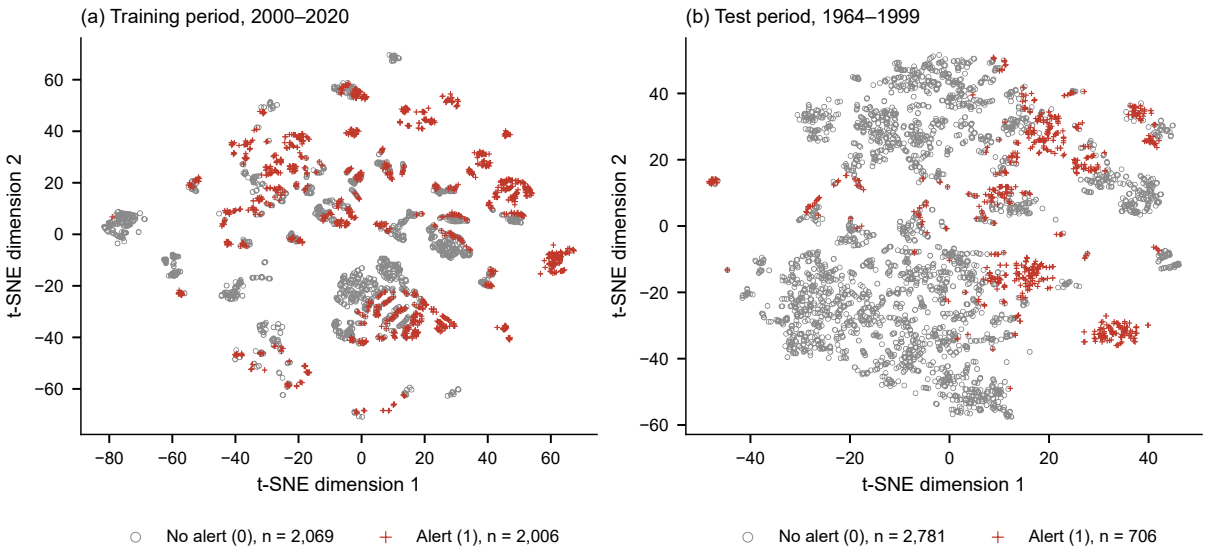

Fig. 3 Two-dimensional t-SNE embedding of the twelve z-scored catalogue and instrument attributes of Experiment 1 (MATLAB `tsne`, random seed 72). (a) Training period, 2000–2020; (b) test period, 1964–1999

vertical component and applied a causal second-order Butterworth band-pass filter of 1–10 Hz. An STA/LTA detector (Allen 1978) compares the mean squared filtered acceleration in the latest 0.25 s with that in the preceding 2 s and emits a trigger when the ratio stays at or above 4 for 0.10 s. We checked the detector against reference P-onset intervals picked visually on 12 development traces (six per period, at 100 and 200 samples per second). Only one trigger fell within its reference interval; most fired earlier, on pre-event noise or glitches (Fig. 4).

Twelve traces are too few to characterise the detector, so we also compared it with an automatic, non-causal reference onset computed for every record: a zero-phase filter, a cumulative-energy search for the onset region, an Akaike information criterion pick, and a backward search for the start of a sustained onset. We treated the reference as reliable when its signal-to-noise ratio was at least 10 dB, and evaluated a record only when the reference onset occurred at least 2.75 s after the start of the record, the earliest time at which the detector can trigger. A trigger was counted as correct when it fell between 0.5 s before and 1.0 s after the reference onset. With this tolerance, the reference fell within 8 of the 12 visual intervals, and the detector was correct on 4 of them. On the 1964–1999 records, the detector was correct for 79.7% of the 650 evaluable records of a development subset and 77.3% of the 677 of a validation subset; most errors were early triggers (18.9% and 21.4%). On the 877 evaluable records of 2000–2020, it was correct for only 41.2% and early for 56.7%. Early triggers arise from three sources: near-zero pre-event noise or digital zeros with isolated pulses, which make the long-term average almost vanish; single-sample spikes; and, in Mexico City records, high background noise ahead of weak, emergent P arrivals.

We therefore also built a corrected detector, tuned only on the 1964–1999 development subset. It adds a causal three-sample median filter, which removes single-sample spikes at a delay of one sample; a floor on the long-term average of $0.1\ \max(0.02\ \mathrm{gal}, q)^2$, where q is the quantisation step of the record; a blind interval in which no trigger is accepted until 0.5 s after the STA and LTA windows are complete; and a threshold of 5 sustained for 0.20 s. It was correct for 95.1% and 93.5% of the evaluable development and validation records and for 70.4% of the evaluable 2000–2020 records, where 17.4% of its triggers were early and 11.5% late. The late triggers correspond to weak, emergent P arrivals that a 0.25/2-s STA/LTA does not capture. The detector was also implemented in MATLAB and matched the Python version within one sample on 50 test traces.

Neither detector is a P pick, so in the main analysis we measure window duration from trigger emission: a nominal 3-s window contains the first 3 s of signal after the detector fired, whatever the true onset. The reference evaluation also reveals a domain shift between the periods. All the 2000–2020 records come from stations in Mexico City, against 12% of the 1964–1999 records, and the median epicentral distance is 325 km, against 55 km. With the original detector, 66% of the 3-s windows of the 2000–2020 test sample have a signal-to-noise ratio below 3 dB.

Feature windows start at the sample after the trigger and contain 1, 2 or 3 s of data. We subtracted from each channel its mean over the same 2-s background interval used by the trigger. The vertical configuration uses five features of the vertical acceleration: absolute peak, root mean square, mean absolute value, peak-to-peak range and the integral of squared acceleration. The horizontal configuration uses the absolute peaks and squared-acceleration integrals of the two horizontal components, each pair sorted by value so that the features do not depend on sensor orientation. The three-component configuration combines these nine features with the sum of the squared-acceleration integrals over the three components and the log power ratio $\log\left(1 + P_w / \max(P_b, 10^{-12})\right)$, where P_w and P_b are the mean squared accelerations pooled across channels in the window and in the background interval. All features are in gal (gal^2 s for integrals). These amplitude and energy measurements are close in spirit to the inputs of the $(t_S - t_P)$ and $(t_P + 3\,\mathrm{s})$ algorithms, but they are computed from a single station and without station calibration.

4.3 Sample selection

We required the same sampling frequency in all three channels, an identifiable vertical component, two horizontal components separated by 90° ± 5°, and non-constant signal in all three components over the full 3-s window. The trigger audit recorded 6,723 triggers and 522 evaluated records without a trigger; other records were set aside because of label, waveform or sampling-metadata issues, including a declared incomplete start. Of the 6,675 records with all three post-trigger windows, 14 had origin dates outside their nominal archive period and 13 failed the three-component criteria. The final common sample contains 2,630 training records from 1964–1999 (509 positive) and 4,018 test records from 2000–2020 (1,977 positive). All feature configurations and window durations use exactly these records, so their differences are paired. With the corrected labels of Section 3.1, the same sample has 508 positive training records and 2,070 positive test records: 93 test labels change from no alert to alert, and one training label changes from alert to no alert through the event-level assignment.

For the robustness analysis of Section 5.5 we built two further samples, both with the corrected labels and with the same quality and three-component criteria. Each of them is evaluated with two sets of windows on exactly the same records, so the effect of the window start is paired. The first sample contains the records with usable windows from both the original and the corrected detectors: 2,192 training records (373 positive) and 3,489 test records (1,912 positive). The corrected detector did not trigger on 953 records that are usable with the original detector; for 617 of them, the automatic reference places the onset before 2.75 s, when the corrected detector cannot trigger. The second sample contains the records with a reliable reference onset at or after 2.75 s and usable windows from both the reference onset and the original detector: 1,305 training records (150 positive) and 863 test records (650 positive). In this sample, windows start at the reference onset sample, and the background interval is the 2 s that end 0.25 s before it, as for the trigger.

4.4 Models and chronological evaluation

Training on the older period and testing on the later one mimics the use of an alert model on future earthquakes. We normalised event dates and origin times from the headers and verified that the two subsets do not overlap in time. The normalised event date and origin time define 189 groups in the test set, which serve as approximate event identifiers.

For the matched feature comparison we fitted a separate model for every configuration and window with two learners: the legacy tree, kept with its original class coding, and adaptive boosting (MATLAB AdaBoostM1 with 1,000 learning cycles and random seed 72) (Freund and Schapire

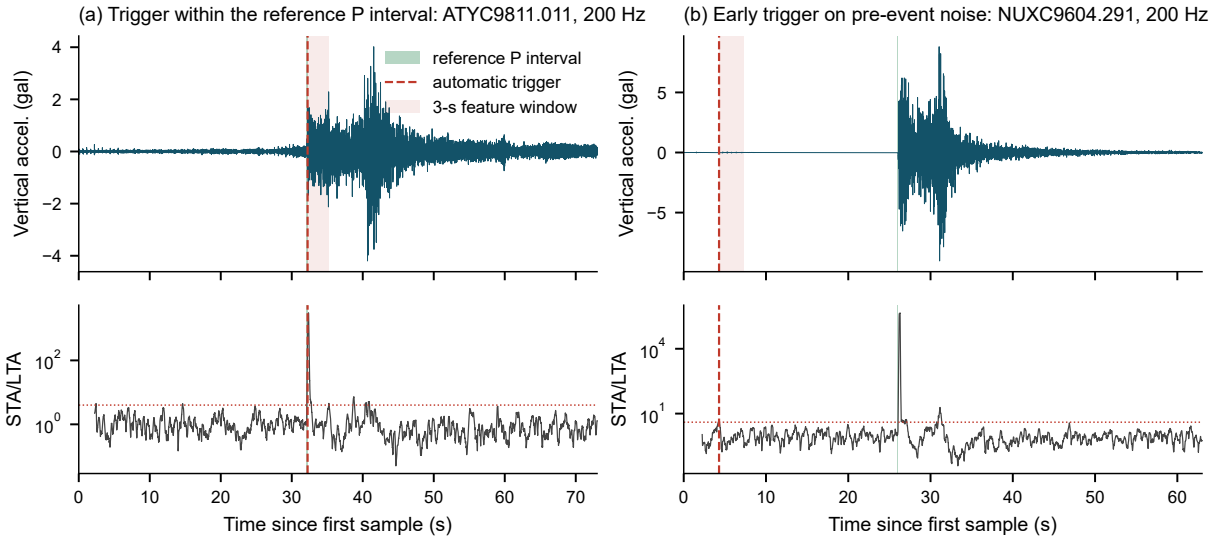

Fig. 4 Automatic trigger on two development traces (vertical component). Top: acceleration with the reference P-onset interval (shaded), the trigger (dashed) and the 3-s feature window. Bottom: STA/LTA ratio of the 1–10 Hz filtered signal and the threshold of 4 (dotted). (a) The trigger falls within the reference interval. (b) The trigger fires on pre-event noise about 22 s before the P onset, so the feature window contains no earthquake signal

1997). To check that conclusions about the vertical features do not depend on the learner, we also screened 19 classifiers on the vertical features, using a closely matched sample selected before the three-component criteria were applied (2,641 training and 4,020 test records). The screen includes the legacy tree with labels recoded as 1 and 2, which removes its indexing inconsistency, and an established tree implementation (MATLAB `fitctree`, CART), as well as the classifiers of Experiment 1, bagged trees, support vector machines with linear and radial-basis kernels, a multilayer perceptron and two recurrent networks applied to the feature vector (Table 8).

4.5 Performance measures and uncertainty

The unit of evaluation is the accelerogram. With TP, FN, FP and TN the counts of true positives, false negatives, false positives and true negatives for the alert class, we report accuracy, precision $P = TP/(TP + FP)$, sensitivity $R = TP/(TP + FN)$, the false-alarm rate $\mathrm{FPR} = FP/(FP + TN)$ and balanced accuracy $\mathrm{BA} = (R + 1 - \mathrm{FPR})/2$ (Witten et al. 2011; He and Garcia 2009). Balanced accuracy equals 50% for any constant or random predictor regardless of class balance, whereas accuracy does not: on the 1964–1999 records, predicting “no alert” for every record gives 79.75% accuracy. Missed alerts (FN) and false alarms (FP) have different social costs, so we report both. All models produce hard labels; for a hard-label classifier the area under the ROC curve equals balanced accuracy (Fawcett 2006), so we show operating points rather than curves and do not report AUC separately.

For the difference in balanced accuracy between the three-component and vertical configurations, we computed paired 95% percentile intervals from 10,000 bootstrap resamples of the 189 test event groups. These intervals describe uncertainty due to the composition of the test set, conditional on the fitted models, and account for the correlation between records of the same earthquake. The analysis is exploratory: the 2000–2020 records were examined during the original study, so they are not a pristine holdout.

5 Results

5.1 Experiment 1: cross-validation on the training period

Table 5 shows the out-of-fold accuracies on the 2000–2020 records. The legacy tree classified every held-out fold without error in all ten repetitions, because it converged to the same tree in every fold.

Table 5 Experiment 1: ten-fold cross-validation accuracy (%) on the 2000–2020 catalogue attributes, for the best of ten repetitions (out-of-fold estimates)

Classifier	Mean	Median	Min	Max
Legacy tree	100.00	100.00	100.00	100.00
Naive Bayes	99.68	99.75	98.77	100.00
k-NN ($k = 3$)	97.96	98.41	96.81	98.77
BP-ANN	85.15[1]	83.31	81.08	87.72
Decision tables	75.71	76.44	69.85	80.15
Legacy tree, 1964–2020[2]	99.97	100.00	99.87	100.00

Values from the archived cross-validation outputs of the original study.

[1]Mean over all repetitions; median, minimum and maximum of the best repetition.

[2]Cross-validated on the merged 1964–2020 matrix.

Naive Bayes (mean 99.68%) and k-NN (97.96%) were nearly as accurate. For k-NN, the mean accuracy decreased monotonically from 98.86% at $k = 3$ to about 94.3% at $k = 21$, consistent with the small, interleaved islands of Fig. 3a: the larger the neighbourhood, the more often it includes records of the other class. BP-ANN (85.15%) and decision tables (75.71%) varied more between folds.

5.2 Experiment 1: transfer to the older period

Applying Eq. 5 directly to the archived matrices reproduced every historical label, without training. Table 6 shows the classifiers of Experiment 1 on the 1964–1999 test period. The saved legacy tree reproduced the confusion counts reported in the original study: it missed 18 of 706 positive records and produced 3 false alarms among 2,781 negatives (balanced accuracy 98.67%, lift 4.92 against a maximum of 4.94). BP-ANN and decision tables also transferred well (95.10% and 95.05%), although decision tables traded fewer missed alerts (28 against 55) for three times as many false alarms (165 against 56). Naive Bayes and k-NN, which scored above 97% out of fold on the training period, dropped to balanced accuracies of 69.44% and 61.12%.

The toolbox classifiers (Table 7) confirm that accuracy alone is misleading on the imbalanced test period. Only logistic regression, SGD and LDA exceeded the 79.75% of a constant "no alert" predictor by a wide margin; the random forest barely exceeded it, and AdaBoostM1, at 51.28%, was well below it.

Figure 5 explains the success of the tree. It uses only three of the twelve attributes: magnitude, distance and focal depth. Its main splits are magnitude thresholds of 5.3, 5.5 and 6.0 and a distance threshold of 346.68 km, close to the 350 km of Eq. 5; focal depth refines decisions near the lower magnitude thresholds, which absorbs the mismatch between the training data and the 5.0 threshold of the rule. The tree has rediscovered the rule that generated the labels. Consistently, when cross-validated on the merged 1964–2020 matrix it reached a mean accuracy of 99.97% (Table 5).

5.3 Experiment 2: vertical waveform features

When the catalogue attributes are replaced by vertical-component features from the first seconds after the trigger, performance falls to near chance (Table 8). Across the 19 classifiers and the three windows, balanced accuracy ranged from 49.37% to 52.50%. Several linear models and neural networks predicted no or almost no positive records, which gives a balanced accuracy of 50% by construction. The tree learners, k-NN and the ensembles flagged more records, but with similar proportions of positive and negative labels. Recoding the labels of the legacy tree (51.66% at 3 s) and replacing it with CART (51.40%) did not change this picture, so the indexing inconsistency of the legacy tree is not the reason for its low scores.

5.4 Experiment 2: horizontal and three-component features

Table 9 compares the feature configurations on the common sample, with the historical labels. Horizontal features alone gave 49.21% to 51.46%, no better than the vertical features. The three-component configuration outperformed the vertical configuration in all six matched runs. At 3 s, it increased balanced accuracy by 4.90 percentage points for AdaBoostM1 (95% bootstrap interval 2.74–7.14) and by 3.85 points for the legacy tree (1.42–6.33). The gain for AdaBoostM1 was also clear at 1 s (2.95; 1.11–4.71) and 2 s (4.66; 2.62–6.61), whereas the intervals for the legacy tree at 1 and 2 s included zero (Fig. 7).

The absolute performance remained low. At 3 s, the three-component AdaBoostM1 model recovered 484 of 1,977 positive labels (sensitivity

Table 6 Experiment 1: classifiers trained on 2000–2020 catalogue attributes and tested on the 1964–1999 records (706 positive, 2,781 negative). [a]Confusion counts reproduced in the present audit from the saved model. [b]Counts derived from the accuracy, precision and sensitivity reported in the original study; the producing runs were not re-executed

Model	TP	FN	FP	TN	Accuracy (%)	P (%)	R (%)	FPR (%)	BA (%)
Eq. 5 applied directly	706	0	0	2,781	100.00	100.00	100.00	0.00	100.00
Legacy tree[a]	688	18	3	2,778	99.40	99.57	97.45	0.11	98.67
BP-ANN[b]	651	55	56	2,725	96.82	92.08	92.21	2.01	95.10
Decision tables[b]	678	28	165	2,616	94.47	80.43	96.03	5.93	95.05
Naive Bayes[b]	456	250	715	2,066	72.33	38.94	64.59	25.71	69.44
k-NN ($k = 3$)[b]	423	283	1,048	1,733	61.83	28.76	59.92	37.68	61.12

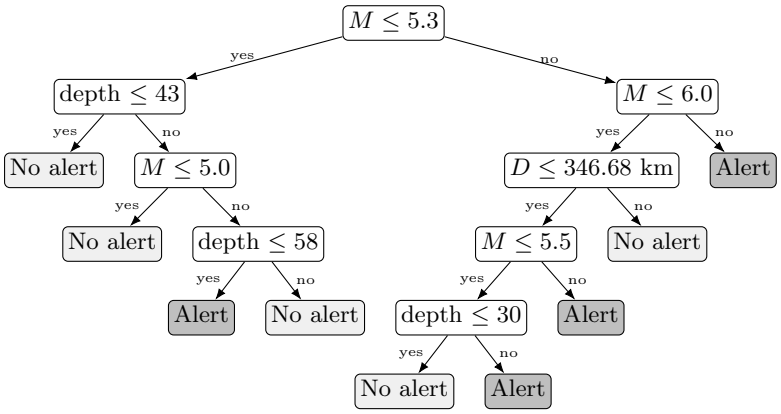

Fig. 5 Decision tree learned by the legacy tree from the 2000–2020 catalogue attributes (identical in every cross-validation fold). M: catalogue magnitude; D: distance to the reference point; depth: focal depth in km. Dark leaves: alert

Table 7 Experiment 1: toolbox classifiers with default settings, tested on 1964–1999 (accuracy reported in the original study; confusion matrices were not archived). A constant "no alert" predictor reaches 79.75%

Classifier	Accuracy (%)
Logistic regression	94.18
Linear model, SGD	93.83
Linear discriminant analysis	92.08
Support vector machine	86.05
Random forest	82.54
AdaBoostM1	51.28

24.48%) and flagged 275 of 2,041 negative labels (false-alarm rate 13.47%); its vertical counterpart recovered 292 positives and flagged 277 negatives. The three-component legacy tree recovered 512 positives and flagged 358 negatives, against 481 and 483 for its vertical counterpart. The gain from adding horizontal information thus came mainly from more true positives for AdaBoostM1 and from fewer false alarms for the tree. Figure 6 places all models of both experiments in ROC space: the catalogue-attribute models lie near the top-left corner, whereas every waveform model lies close to the diagonal of random decisions.

5.5 Experiment 2: corrected labels and window start

Table 10 repeats the matched comparison of Table 9 with the same learners, settings and random seed, in MATLAB R2021a. With the historical labels, the rerun reproduced all twelve balanced accuracies of Table 9. The intervals in this section are paired 95% percentile intervals computed as in Section 4.5, from 10,000 bootstrap resamples of the test event groups (189 in the common sample, 177 in Sample 1 and 94 in Sample 2).

Correcting the labels changed the twelve balanced accuracies by −0.61 to +1.36 points. For three-component AdaBoostM1 at 3 s, balanced accuracy went from 55.50% to 54.89% (difference

Table 8 Experiment 2, vertical-feature screen: balanced accuracy (%) on the 2000–2020 test records (4,020; 1,978 positive) for models trained on 1964–1999, by window after the trigger, and positive predictions at 3 s

Classifier	1 s	2 s	3 s	TP_{3s}	FP_{3s}
k-NN ($k = 3$)	51.09	51.12	52.04	381	310
Random forest	51.69	50.63	51.77	287	224
Bagged trees	51.75	50.43	51.70	258	197
Legacy tree (1/2)	51.45	51.97	51.66	519	468
AdaBoostM1	52.50	50.47	51.41	264	215
CART (`fitctree`)	51.58	49.37	51.40	435	392
GRU	50.08	50.29	51.07	82	41
Decision tables	50.20	51.11	50.76	30	0
SVM, radial basis	51.12	50.55	50.65	47	22
Legacy tree (0/1)	52.21	50.40	50.49	495	491
MLP (100–50)	50.28	50.19	50.26	25	15
LSTM	50.03	50.10	50.18	14	7
Logistic regression	50.00	50.00	50.10	5	1
LDA	50.00	50.00	50.03	1	0
Linear SVM	50.00	50.00	50.00	0	0
SGD logistic	50.00	50.00	50.00	0	0
Legacy tree, PRF	50.66	50.41	50.00	0	0
BP-ANN	50.00	50.00	50.00	0	0
Naive Bayes	49.89	49.72	49.95	25	28

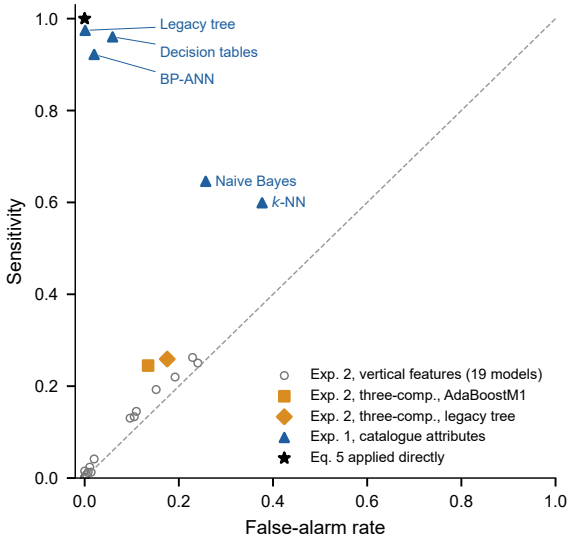

Fig. 6 Operating points in ROC space. Experiment 1 (triangles): catalogue-attribute models tested on 1964–1999. Experiment 2 at 3 s (circles, square, diamond): waveform-feature models tested on 2000–2020. The dashed diagonal corresponds to random decisions. The test periods differ between experiments

Table 9 Experiment 2: balanced accuracy (%) by model, feature configuration and window after the automatic trigger, on the same 4,018 test accelerograms (2000–2020)

Model	Features	1 s	2 s	3 s
Legacy tree	Vertical	52.07	51.05	50.33
	Horizontal	49.21	51.46	49.35
	Three-comp.	53.26	52.33	54.18
AdaBoostM1	Vertical	52.11	50.16	50.60
	Horizontal	50.53	49.95	50.57
	Three-comp.	55.06	54.82	55.50

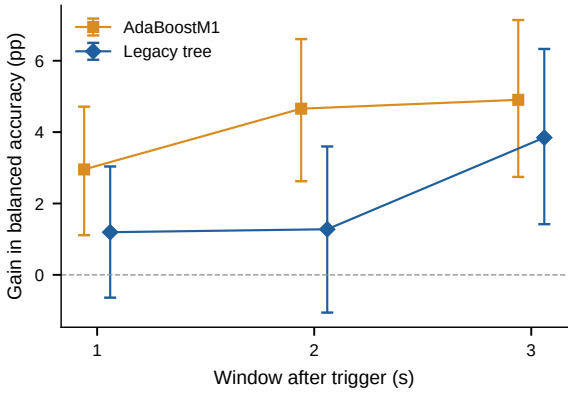

Fig. 7 Paired gain in balanced accuracy of the three-component over the vertical feature configuration. Bars show 95% percentile intervals from 10,000 bootstrap resamples of the 189 test event groups

−0.61; −1.68 to 0.68): the model recovered 450 of 2,070 positive labels (sensitivity 21.74%) with a false-alarm rate of 11.96%. The three-component configuration still outperformed the vertical configuration in all six matched runs. At 3 s, the gain was 3.09 points for AdaBoostM1 (1.04–5.22) and 4.34 for the legacy tree (1.82–7.02). The gain for AdaBoostM1 remained clear at 1 s (2.66; 0.78–4.41) and 2 s (4.02; 2.06–6.03), whereas the intervals for the legacy tree at 1 and 2 s included zero.

On the first robustness sample (Section 4.3), windows from the corrected detector raised the balanced accuracy of three-component AdaBoostM1 to 58.72%, 57.33% and 57.44% at 1, 2 and 3 s, against 56.18%, 55.90% and 55.11% with the original detector on the same records. None of these paired differences was distinguishable from zero: 2.54 points (−0.41 to 5.53), 1.43 (−2.01 to 4.71) and 2.34 (−0.75 to 5.45). The other configurations changed in both directions. The vertical legacy tree lost 2.84 points at 1 s (−5.53 to −0.19), vertical AdaBoostM1 gained

2.66 points at 3 s (0.25–5.19), and the intervals of the remaining runs included zero. At 3 s, three-component AdaBoostM1 recovered 646 of 1,912 positive labels (sensitivity 33.79%) with a false-alarm rate of 18.90%, against 547 (28.61%) and 18.39% with the original detector.

On the second robustness sample, windows that start at the reference onset did not improve classification. Balanced accuracies ranged from 45.66% to 55.38%, and none of the twelve paired differences with windows from the original detector on the same records was distinguishable from zero. This sample is small and imbalanced: its 863 test records come from 94 event groups and 75% of them are positive, so these estimates are imprecise. Across all samples, label versions and window starts, balanced accuracy remained between 45.66% and 58.72%. Correcting the labels or the window start therefore did not change the conclusions of Sections 5.3 and 5.4.

6 Discussion

6.1 What catalogue-attribute scores measure

The two experiments answer different questions and should not be compared as if they measured the same skill. Experiment 1 asks whether a classifier can rediscover a threshold rule from the variables that define it; the answer is yes for a decision tree, and the rule itself does so exactly. The near-perfect score of the tree, which was the main result of the original study, is therefore an upper bound that no early-warning algorithm can approach, because the catalogue magnitude is known only after the event.

Experiment 1 does, however, carry a useful negative result about temporal transfer. Naive Bayes and k-NN were almost perfect out of fold within the 2000–2020 period yet lost 30 to 37 points of balanced accuracy on 1964–1999. Both rely on exact or nearest training values of all twelve attributes. The periods differ in their magnitude and depth distributions and in some instrument values (Fig. 1), and the combinations of attributes form period-specific islands (Fig. 3), so the nearest or most frequent training values of one period are poor guides to the other. The tree, in contrast, relies on the two attributes that define the label. Cross-validation within one period therefore says little about performance in another, the behaviour expected under distribution shift (Gama et al. 2014), and alert models trained on archives should be evaluated across time.

6.2 Information in early single-station waveforms

Experiment 2 removes the catalogue inputs and shows how little of the retrospective decision is recoverable from simple amplitude and energy measurements at one station in the first seconds after a trigger. The low scores are consistent with the physics of the problem. A single station's early amplitude depends on magnitude, but also on hypocentral distance, site response, radiation pattern and instrument, and the label additionally depends on the distance from the epicentre to Mexico City, which a single trace does not reveal.

The robustness analysis (Section 5.5) indicates that neither label errors nor trigger errors account for the low scores. Correcting the labels changed balanced accuracy by at most 1.4 points. The corrected detector raised the proportion of correct triggers on the evaluable 2000–2020 records from 41.2% to 70.4%, yet it raised the balanced accuracy of three-component AdaBoostM1 by only 1.4 to 2.5 points, none of these gains was distinguishable from zero, and the model still recovered only about a third of the alert records. Windows aligned with the reference onset did not improve classification either, although that sample is small. Trigger errors still matter for individual records: when the detector fires on pre-event noise, as in Fig. 4b, the feature window contains no earthquake signal at all. In aggregate, however, the main limitation appears to lie in the recording conditions of the test period. All 2000–2020 records come from stations in Mexico City, at a median epicentral distance of 325 km, and two thirds of their 3-s windows have a signal-to-noise ratio below 3 dB (Section 4.2). At such distances and noise levels, the first seconds of the P wave plausibly carry too little energy above the urban background to separate events above and below the alert threshold. The training records, of which 88% come from stations outside Mexico City at a median distance of 55 km, cover different conditions, so the models also face a domain shift between the periods. Both effects separate this

Table 10 Experiment 2, robustness analysis: balanced accuracy (%) on the 2000–2020 test records for models trained on 1964–1999, by learner, feature configuration and window. Common sample: the 2,630 training and 4,018 test records of Table 9, with the historical and the corrected labels. Sample 1: 2,192 training and 3,489 test records, windows from the original and the corrected detectors. Sample 2: 1,305 training and 863 test records with a reliable reference onset, windows from the original detector and from the reference onset. Samples 1 and 2 use the corrected labels (Section 4.3)

			Common sample		Sample 1		Sample 2	
Model	Features	Window (s)	Historical labels	Corrected labels	Original detector	Corrected detector	Original detector	Reference onset
Legacy tree	Vertical	1	52.07	52.47	52.60	49.75	48.34	50.30
		2	51.05	51.59	52.43	50.22	51.27	48.01
		3	50.33	50.95	51.15	52.02	48.62	45.66
	Three-comp.	1	53.26	54.25	54.56	51.84	51.16	50.60
		2	52.33	53.62	53.73	55.19	53.60	50.82
		3	54.18	55.30	53.99	52.08	51.00	50.52
AdaBoostM1	Vertical	1	52.11	52.46	52.32	51.32	49.68	49.29
		2	50.16	51.52	51.50	52.79	51.78	46.33
		3	50.60	51.80	51.59	54.25	49.92	48.77
	Three-comp.	1	55.06	55.12	56.18	58.72	52.66	55.38
		2	54.82	55.55	55.90	57.33	50.53	49.00
		3	55.50	54.89	55.11	57.44	50.53	47.76

setting from that of SASMEX, which relies on near-source stations.

The SASMEX algorithms address part of the remaining ambiguity with station-specific calibration and by requiring two stations to agree, neither of which our features include. The gain from horizontal and pooled energy features was consistent: it held in all matched runs and persisted with the corrected labels (3.09 points for AdaBoostM1 at 3 s). This agrees with the use of both vertical and horizontal energy by the $2(t_S - t_P)$ algorithm, but the three-component configuration added horizontal peaks, pooled energy and the background-relative power ratio together, so their separate contributions remain unresolved. We did not estimate the uncertainty of differences between window durations.

6.3 Relation to the SASMEX algorithms

The original study reimplemented the three SASMEX algorithms in MATLAB from their publications, with the published regression coefficients, and applied them to every archived record. The reimplementation illustrates how the algorithms behave on individual records. On record AU4620090427164627 (M 5.7), the detected P and S arrivals were 28.93 s apart, beyond the 24-s search window for the S wave, so no algorithm produced an estimate and none issued an alert; a station closer to the epicentre would have been needed. On record C16620170919181439, of the 2017 Puebla–Morelos earthquake (M_w 7.1), the S–P time was 21.92 s; the $2(t_S - t_P)$ algorithm classified the event as $M \geq 6.0$ and the $(t_S - t_P)$ algorithm estimated $M = 6.5$, whereas the $(t_P + 3\,\mathrm{s})$ algorithm returned no estimate because $a_{v(T)}$ fell outside the ranges of its piecewise regression.

The aggregate scores of that reimplementation (Appendix **??**) were close to chance on 1964–1999, but we do not use them to compare SASMEX with the classifiers, for four reasons. First, the station coefficients α_k and β_k of the P and S detector and their calibration method are not published, and the reimplementation drew them from a uniform distribution on $(0, 1)$. Second, the publications assume 200 samples per second, whereas part of the archive was recorded at 100, and some files contain a zero vertical channel. Third, the $(t_S - t_P)$ and $(t_P + 3\,\mathrm{s})$ algorithms estimate magnitude only, whereas Eq. 5 also depends on distance, so even perfect magnitude estimates would disagree with some labels. Fourth, the archived comparison script modifies the saved prediction vectors at random before scoring, and direct evaluation of those vectors did not reproduce the reported tables. A fair comparison requires station-calibrated coefficients and, ideally, operational magnitude estimates and alert logs, which we plan to request from the developers of SASMEX.

6.4 Limitations

The labels are retrospective and derive from catalogue magnitudes of mixed scales; using the

maximum instead of the first listed magnitude would change 182 labels. Reading errors of the legacy extractor affected 101 labels, all in the test period, and four older records without earthquake data remain unlabelled. Experiment 1 keeps the historical labels to stay comparable with the original study, so its scores refer to the rule as originally applied; in Experiment 2 the corrected labels changed balanced accuracy by at most 1.4 points. Both detectors are simple STA/LTA triggers. The original detector fired early on more than half of the evaluable test records, and the corrected detector, although more accurate, cannot trigger in the first 2.75 s of a record and misses weak, emergent onsets. The 12 visual reference intervals were picked by the authors with the assistance of a language model (Section **??**), not by a seismic analyst. The reference onset used to evaluate the detectors is an automatic, non-causal pick and is uncertain for distant records with low signal-to-noise ratio. Windows aligned with that onset could be evaluated only on a small and imbalanced subset (94 test event groups), so the effect of exact onset alignment remains poorly constrained, and onsets that occurred before the start of a record are not represented. All test records come from stations in Mexico City, at a median epicentral distance of 325 km, and most of their early windows have a low signal-to-noise ratio; the results therefore describe distant urban stations rather than the near-source stations on which SASMEX relies. The sample excludes records without usable three-component windows, which may favour some stations or periods. The two periods differ in instrumentation, station coverage and class balance; a single forward split cannot separate these effects, and rolling or expanding training windows would give a more complete picture of temporal transfer. The evaluation unit is the accelerogram, not the event or the alert centre. Finally, the study is offline: a prospective evaluation would need to quantify triggers and false alarms on continuous streams, the time from origin to decision, and the warning time available at the target.

7 Conclusions

We audited an archive of 7,562 Mexican accelerograms from 1964–2020 labelled with a retrospective version of the SASMEX dissemination criteria and evaluated two ways of recovering those labels. Classifiers given catalogue magnitude and distance largely reproduce the labelling rule, so their near-perfect scores, including the 99.40% of the legacy tree on the older period, measure rule recovery rather than early-warning skill; the same experiment shows that near-perfect cross-validation within one period does not guarantee transfer to another. When only acceleration from the first 1–3 s after an automatic trigger was available, 19 classifiers using vertical features performed near chance under a forward chronological split, and three-component features raised balanced accuracy by about 4–5 percentage points at 3 s, to at most 55.50%. Even then, about three quarters of alert-worthy records were missed.

These conclusions survived the corrections made during the audit. Reading errors of the legacy extractor had changed 101 labels; with the corrected labels, balanced accuracy changed by at most 1.4 points and the three-component gain remained (3.09 points at 3 s for adaptive boosting; 95% interval 1.04–5.22). A corrected detector raised the share of correct triggers on the evaluable test records from 41.2% to 70.4%, but its gains in balanced accuracy, at most 2.54 points for the three-component model, were not distinguishable from zero. Windows that start at a reference P onset did not perform better than windows from the original trigger. Across all these variants, balanced accuracy stayed between 45.66% and 58.72%, so neither label nor trigger errors explain the low skill of single-station early windows in this archive.

Future work should use more reliable P pickers (noise-relative, adaptive or learned thresholds) validated against analyst picks, combine several stations as in operational EEW, define targets that reflect the information available at decision time, evaluate with rolling chronological windows and per-event measures, and seek station-calibrated SASMEX outputs for a matched comparison. Other open archives with analyst picks, such as those of the IRIS Data Management Center and the Southern California Earthquake Data

Center, would allow the trigger and the features to be validated independently.

Acknowledgements. We thank the Instituto de Ingeniería, UNAM, for access to the RAII-UNAM strong-motion records and for permission to use them and to publish the data derived from them.

Statements and Declarations

Funding María del Carmen Ríos Perea received a master's scholarship from SECIHTI (formerly CONAHCYT), CVU 1142000.

Competing interests The authors declare no competing interests.

Author contributions M.C.R.P.: conceptualisation, data curation, software, investigation, writing (original draft). B.M.M.: conceptualisation, methodology, supervision, formal analysis, writing (review and editing). R.M.R.: supervision, methodology, writing (review and editing).

Data availability The original accelerograms used in this study, 7,562 ASA 2.0 records from 1964–2020, are available from the strong-motion database of the Instituto de Ingeniería, UNAM (Instituto de Ingeniería, UNAM 2023); they were used with the permission of the Instituto de Ingeniería, UNAM, and are not redistributed here.

The data derived in this study are available on GitHub (https://github.com/opelo1209/SASTREMORS) and archived at Zenodo (https://doi.org/10.5281/zenodo.24215086). They comprise the tre-data 1.0 feature matrices with the historical labels; the attributes recomputed from the headers, the record-to-matrix correspondence and the corrected labels; the reference onsets and the triggers of the original and corrected detectors; the Experiment 2 features for each sample, window and feature configuration; and the record-level predictions, performance metrics and bootstrap intervals. Every table identifies records by their RAII-UNAM file names, so it can be linked to the original accelerograms. The derived data contain no personal information.

Code availability The Python and MATLAB scripts used for the audit, triggering, feature extraction, classification, bootstrap analysis and figures are available on GitHub (https://github.com/opelo1209/SASTREMORS) and archived at Zenodo (https://doi.org/10.5281/zenodo.24215086).

References

Allen RM, Melgar D (2019) Earthquake early warning: Advances, scientific challenges, and societal needs. Annual Review of Earth and Planetary Sciences 47:361–388. https://doi.org/10.1146/annurev-earth-053018-060457

Allen RV (1978) Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America 68(5):1521–1532

Breiman L (2001) Random forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20:273–297. https://doi.org/10.1007/BF00994018

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964

Cuéllar A, Suárez G, Espinosa-Aranda JM (2017) Performance evaluation of the earthquake detection and classification algorithm 2(tS–tP) of the Seismic Alert System of Mexico (SASMEX). Bulletin of the Seismological Society of America 107(3):1451–1463. https://doi.org/10.1785/0120150330

Cuéllar A, Suárez G, Espinosa-Aranda JM (2018a) An earthquake early warning algorithm based on the P-wave energy released in the tS–tP interval. Bulletin of the Seismological Society of America 108(1):238–247. https://doi.org/10.1785/0120170115

Cuéllar A, Suárez G, Espinosa-Aranda JM (2018b) A fast earthquake early warning algorithm based on the first 3 s of the P-wave coda. Bulletin of the Seismological Society of America 108(4):2068–2079. https://doi.org/10.1785/0120180079

Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29:103–130. https://doi.org/10.1023/A:1007413511361

Espinosa-Aranda JM, Jiménez A, Contreras O, et al (1992) Mexico city seismic alert system. In: Proc. International Symposium on Earthquake Disaster Prevention. CENAPRED-JICA, Mexico City, Mexico, pp 315–324

Espinosa-Aranda JM, Cuéllar A, García A, et al (2009) Evolution of the Mexican Seismic Alert System (SASMEX). Seismological Research Letters 80(5):694–706. https://doi.org/10.1785/gssrl.80.5.694

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1):119–139. https://doi.org/10.1006/jcss.1997.1504

Gama J, Žliobaitė I, Bifet A, et al (2014) A survey on concept drift adaptation. ACM Computing Surveys 46(4):44. https://doi.org/10.1145/2523813

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

Instituto de Ingeniería, UNAM (2023) Base de datos de registros acelerográficos de la Red Sísmica Mexicana. https://aplicaciones.iingen.unam.mx/AcelerogramasRSM/Inicio.aspx

Kanamori H (2005) Real-time seismology and earthquake damage mitigation. Annual Review of Earth and Planetary Sciences 33:195–214. https://doi.org/10.1146/annurev.earth.33.092203.122626

Kohavi R (1995) The power of decision tables. In: Proc. 8th European Conference on Machine Learning (ECML), Lecture Notes in Computer Science, vol 912. Springer, pp 174–189, https://doi.org/10.1007/3-540-59286-5_57

Kong Q, Allen RM, Schreier L, et al (2016) MyShake: A smartphone seismic network for earthquake early warning and beyond. Science Advances 2(2):e1501055. https://doi.org/10.1126/sciadv.1501055

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. Journal of Machine Learning Research 9:2579–2605

Mousavi SM, Ellsworth WL, Zhu W, et al (2020) Earthquake transformer: An attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Communications 11:3952. https://doi.org/10.1038/s41467-020-17591-w

Münchmeyer J, Bindi D, Leser U, et al (2021) The transformer earthquake alerting model: A new versatile approach to earthquake early warning. Geophysical Journal International 225(1):646–656. https://doi.org/10.1093/gji/ggaa609

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Sarker IH (2021) Machine learning: Algorithms, real-world applications and research directions. SN Computer Science 2(3):160. https://doi.org/10.1007/s42979-021-00592-x

Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, Burlington, MA