Fulfilling the information need after an earthquake. Statistical modelling of citizen science seismic reports for predicting earthquake parameters in near real time

Francesco Finazzi

Department of Management, Information and Production Engineering, University of Bergamo, Dalmine, Italy.

E-mail: francesco.finazzi@unibg.it

Summary. When an earthquake affects an inhabited area, a need for information immediately arises among the population. In general, this need is not immediately fulfilled by official channels which usually release expert validated information with delays of many minutes.
Seismology is among the research fields where citizen science projects succeeded in collecting useful scientific information. More recently, smartphone ubiquity is giving the opportunity to involve even more citizens.
This paper focuses on seismic intensity reports collected through smartphone applications while an earthquake is occurring. The aim is to provide a framework for predicting and updating in near real time earthquake parameters useful for assessing the impact of the earthquake. This is done using a multivariate space-time model based on time-varying coefficients and a spatial latent variable.
As a case study, the model is applied to more than 200,000 seismic reports globally collected over a period of around 4 years by the Earthquake Network citizen science project. It is shown how the time-varying coefficients are needed to adapt the model to an information content that changes with time, and how the spatial latent variable is able to capture the local seismicity and the heterogeneity in the people response across the globe.

Keywords: smartphone app; spatio-temporal modelling; EM algorithm; dynamic kriging; D-STEM software

1. Introduction

Whenever a natural event or disaster affects an inhabited area, an instant need for information arises among the population. This is especially true in a fully connected word where people have immediate access to multiple sources of information (TV, Web, social networks, smartphone applications) and they expect to find answers. In one way or another, this need for information must be fulfilled. If not, people might be susceptible to take the wrong action, either during an emergency or when an emergency does not exist and no action should be taken.

Among all natural events, earthquakes are those posing one of the highest risk for the population. According to the Significant Earthquake Database maintained by the National Centers For Environmental Information of the National Oceanic and Atmospheric Administration, around 570,000 people died and around one million people were injured worldwide due to earthquakes only in the period 2000-2018. Moreover, the seismic risk

map recently produced by the Global Earthquake Model foundation (Silva et al., 2018) clearly shows that very few countries are exempt from potential losses connected to earthquakes.

In seismic areas, it is thus essential to have tools of risk mitigation able to collect and communicate useful information before (Jordan et al., 2011), during (Allen and Kanamori, 2003) and after (Gehl et al., 2018) an earthquake.

From the physical point of view, an earthquake is a sudden release of energy occurring with the rupture and the slip of a fault. Using networks of seismometers, earthquakes are measured in terms of released energy at the rupture. This information, together with instrumental ground motions, geologically based frequency and amplitude-dependent site corrections and earthquake-rupture models is used to predict the shaking intensity across space (known as ShakeMaps, see Wald et al. (1999)), and then to predict the potential impact on the population (Allen et al., 2009).

Although most seismic countries have advanced monitoring networks able to detect and assess earthquakes in a fast manner, official information are often released to the general public with a delay of several minutes. Usually, this is done to release good quality information eventually validated by experts.

According to the survey detailed in Table 1 and taken by 7,067 people, however, 53% of the respondents would like to receive earthquake information as soon as possible, even if just preliminary information. Additionally, around 74% of all the respondents would like to receive this information within 60 seconds from the moment they feel the earthquake. For around 66% of all the respondent, it is important to understand if the epicentre is close to family/friends who live in a different area. Indeed, a peculiarity of earthquakes is that it is hard, for a person, to understand if the shaking has been stronger somewhere else. In most cases, the need for information is the care for the loved ones.

Another proof that this need for information actually exists lies in the fact that it has been exploited to detect the occurrence of earthquakes. The LastQuake smartphone application (app hereafter) of the European-Mediterranean Seismological Centre (Bossu et al., 2018), for instance, is used to detect events when people feel an earthquake and they open the app looking for information. A similar strategy is adopted in social networks monitoring (Sakaki et al., 2010; Earle et al., 2012; Crooks et al., 2013). When people post comments on Twitter immediately after an earthquake, it is for the need to communicate something and to obtain information from others. However, though it is relatively easy to detect the occurrence of an earthquake from social network monitoring, exploiting the location information content is becoming more difficult. Indeed, only 0.7% of the tweets on Twitter contain geolocation information (Graham et al., 2014) and Twitter has announced† that the ability to tag precise location from tweets is being progressively removed.

Within the same context, Earthquake Network (Finazzi, 2016) is a smartphone-based citizen-science project which aims to release rapid and reliable information to the general public about earthquakes while they are happening. This is done, on the one hand, exploiting the accelerometer sensor on-board each smartphone, and on the other, by collecting and analysing seismic intensity reports sent by smartphone users. The smartphone app, therefore, is both the instrument for collecting information and the instrument to

†https://twitter.com/twittersupport/status/1141039841993355264

receive information in the form of notifications or alerts.

In general, there is a trade-off between rapidity and the quality of what can be communicated. As detailed in (Finazzi and Fassò, 2017), data collected by the accelerometers are analysed in an earthquake early warning setting with a temporal dynamics on the scale of seconds. This allows to provide a very preliminary estimate of the epicentre and to alert nearby cities beforehand. On the other hand, seismic intensity reports require up to minutes to be collected and they are analysed to understand the potential impact of the earthquake on the population. This is similar to what is commonly done with macroseismic intensity questionnaires collected through web sites (Atkinson and Wald, 2007; Tosi et al., 2015). An important difference, however, is that smartphone apps provide a user interface which is immediately available to the user, making the collection of reports faster than ever before. Additionally, apps usually provide high accuracy geolocation information (latitude and longitude of the smartphone location) which are automatically included in the report.

Having as reference the fulfilment of the need for information from the population, the main goal of this paper is to understand, using a statistical approach, if the seismic reports sent while earthquake waves propagate across space are informative enough to predict relevant quantities about the earthquake itself. Quantities of interest are either intrinsic parameters of the earthquake (magnitude, depth, etc.) or parameters related to the interaction between the earthquake and the population (e.g. the distance between the earthquake epicentre and a main city). If such predictions could be obtained within the first minute of the earthquake and with good quality, they would represent useful information complementary to the preliminary information released as early warning by Earthquake Network.

To avoid confusion, it is stated here that the term prediction does not refer to the occurrence of future earthquakes. Instead, what is predicted are the "true" earthquake parameters that will be computed and released by geophysics institutions after many minutes from the beginning of the earthquake.

Predictions obtained from reports may also be useful in underdeveloped and developing countries, where national geophysics institutions may not be able to provide fast information but smartphone penetration is relatively high (see Bossu et al. (2015) for a case study of the 2015 Nepal earthquake sequence). Additionally, any prediction based on seismic reports can be instantly notified to the smartphone users, in a virtuous cycle where each citizen provide a piece of information and has a useful service in return.

With respect to works in literature, the above goal is more ambitious for mainly two reasons. First, it is attempted to predict intrinsic parameters of the earthquake, while, in the past, questionnaires has been used to model the spatial distribution of the macroseismic intensity assuming a known magnitude and hypocentre (De Rubeis et al., 1992; Atkinson and Wald, 2007; Sbarra et al., 2010; Cameletti et al., 2017). Second, it is attempted to provide predictions while the earthquake is in progress and the ground shaking is possibly not yet finished. In particular, the following questions are addressed:

(a) Can seismic reports be used to predict intrinsic parameters of an earthquake?
(b) How reliable are such predictions?
(c) After how many seconds are the predictions reliable?
(d) Do predictions keep improving with time?

Answers will be obtained from the analysis of more than 220 thousand seismic reports collected by the Earthquake Network app within more than 4 years and related to more than $1,500$ earthquakes.

The prediction of earthquake parameters will be obtained adopting a statistical model with input the seismic reports and output the parameters of interest. The model will be trained using the collected reports and having as reference the true earthquake parameters given by the seismic catalog. Once trained, the model is used to predict the parameters of any new earthquake from the reports collected in real time by Earthquake Network.

The rest of the paper is structured as follows. Section 2 introduces the Earthquake Network citizen science project, from which seismic reports are taken. Section 3 describes the structure of the seismic reports and the data transformation applied to extract information from the reports. Section 4 introduces a univariate statistical model for the real time estimation of the earthquake magnitude while Section 5 details a multivariate model for the joint estimation of multiple parameters. A case study involving global data is presented in Section 6. Remarks are given in Section 7 while conclusions in Section 8.

2. The Earthquake Network project

The seismic report analysis detailed in this work is based on reports collected by the Earthquake Network project discussed above. During the last 6 years, around 5 million people from all over the world took part to the project, enabling the collection of reach data sets.

The smartphone app allows users to send a report about any earthquake they experience and the app is designed in such a way that the report is sent as fast as possible. Using buttons, the user can report if the earthquake is mild, strong or very strong. It is observed that reports arrive at the server after few seconds from when the ground shaking begins. This particularly holds for people with the app installed for a long time, who are thus "trained" to send a report very quickly if their smartphone is nearby. In effect, these people become spatially distributed sensors monitoring for earthquakes.

As an example, Figure 2 shows on map the reports received by the server of the Earthquake Network project during a 5.2 magnitude earthquake which occurred on December 16th, 2016 off the coast of Puerto Aldea, Chile. The colour of the marker represents the intensity as perceived by the smartphone user: green for mild, yellow for strong and red for very strong. The first reports are located near the epicentre and most of them are yellow. While time passes and the seismic waves reach more distant cities, new reports are collected. As expected, the higher the distance from the epicentre the lower the reported intensity.

This example suggests that the information content of seismic reports dynamically changes over time while seismic waves propagate. Predictions about relevant quantities of the earthquake can be provided as soon as the first reports arrive at the server and then they are updated while time passes.

Table 1. Results of the survey sent to 105,000 users of the Earthquake Network app. The survey was viewed by 12,300 users and taken by 7,067 users. Area 1 is Italy, Area 2 includes Mexico, Central America and South America while Area 3 is the rest of the world.

Q: When it comes to receive information about an earthquake you prefer	Area 1	Area 2	Area 3	Total
A: To wait for accurate information	461	2670	194	3325(47.0%)
A: Have information as soon as possible even if not accurate	883	2515	344	3742(53.0%)
Q: You just felt an quake. How long are you willing to wait before receiving information on epicentre and magnitude?	Area 1	Area 2	Area 3	Total
A: Maximum 15 seconds	456	1966	216	2638(37.3%)
A: Maximum 30 seconds	161	748	79	988(14.0%)
A: Maximum 1 minute	326	1183	108	1617(22.9%)
A: Maximum 2 minutes	145	457	40	642(09.1%)
A: Maximum 5 minutes	175	602	65	842(11.9%)
A: Maximum 10 minutes	81	229	30	340(04.8%)
Q: You just felt a mild earthquake. What is the main reason why you are interested in knowing epicentre and magnitude as soon as possible?	Area 1	Area 2	Area 3	Total
A: To find out if the epicentre is close to family/friends who live in a different area from mine	784	3604	306	4694(66.4%)
A: To know if it was a dangerous quake	527	1405	203	2135(30.2%)
A: Just out of curiosity	33	176	29	238(03.4%)
Q: You just felt an earthquake. What is the first information channel you turn to for information?	Area 1	Area 2	Area 3	Total
A: TV	114	517	80	711(10.1%)
A: Smartphone apps	952	3431	344	4727(66.9%)
A: Twitter	26	346	17	389(05.5%)
A: Facebook	151	387	30	568(08.0%)
A: Instagram	3	35	7	45(00.6%)
A: Phone call with family/friends	44	180	43	267(03.8%)
A: WhatsApp	54	289	17	360(05.1%)

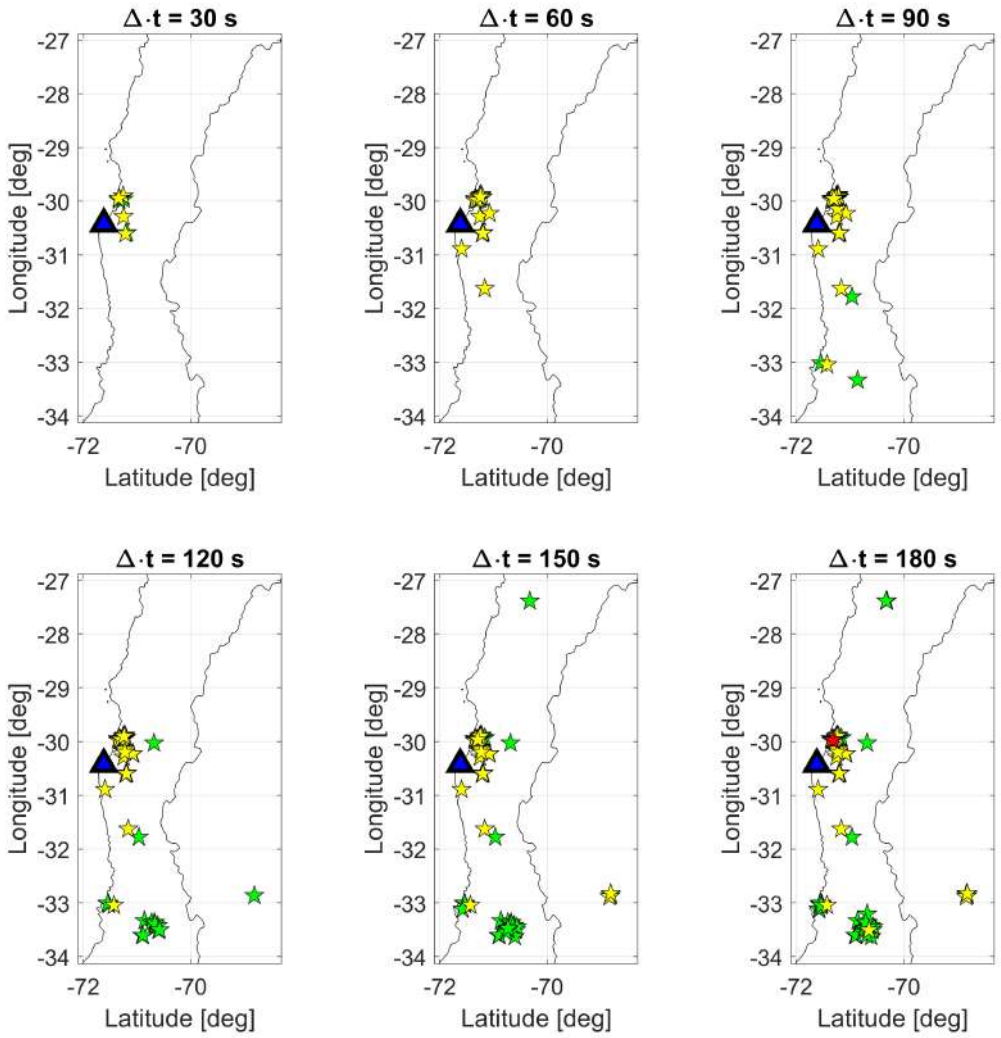

Fig. 1. Spatio-temporal trend of the seismic reports sent by people joining the Earthquake Network citizen science project during a magnitude 5.2 earthquake which occurred on December 16th, 2016 at 23:12:12 UTC off the coast of Puerto Aldea, Chile. The triangular marker is the earthquake epicentre while star markers are the locations of the seismic reports. Green, yellow and red stand for mild, strong and very strong intensity, respectively. From top left to bottom right, the elapsed time from the first report is 30, 60, 90, 120, 150 and 180 seconds.

3. Seismic report data

For a given earthquake felt and reported through the Earthquake Network app, the following data set is available

$$\mathcal{D} = \{(\tilde{\boldsymbol{s}}_j, \tau_j, i_j)\}_{j=1,...,N}$$

where $\tilde{\boldsymbol{s}}_j \in \mathcal{S}^2$ is the spatial location of the smartphone/person over the sphere $\mathcal{S}^2$, τ_j is the report time and $i_j = (green, yellow, red)$ is the earthquake intensity as perceived by the person, with *green* mild, *yellow* strong and *red* very-strong. A posteriori, a total of N reports are collected.

By convention, $\tau_1 = 0$, namely the first report received by the server defines the time origin for the reports. It also follows that the actual origin time of earthquakes is not used nor is it a parameter of interest.

In order to study the dynamics of the reports sent to the server, the discrete time $t = 0, 1, ..., T$ is introduced, with time interval Δ. The actual length of the observation window is then given by $\Delta \cdot T$.

Moreover, $\boldsymbol{s} \in \mathcal{S}^2$ is the centre of gravity of the coordinates of the reports collected at time $t = 1$. The first reports usually come from people living near the epicentre and $\boldsymbol{s}$ computed at $t = 1$ is representative of the area where people are expected to experience strong ground shaking. Spatial location $\boldsymbol{s}$ will be used as reference spatial location when defining statistical models in the following sections. That said, it is stressed that $\boldsymbol{s}$ does not pretend to be an estimate of the epicentre.

In the sequel, for any given $0 \leq t \leq T$, the focus will be on the reports collected in $[0, t]$. In particular, the following quantities will be of interest

$$\begin{aligned} green_0^t &= \frac{\sum_{j=1}^N I\left(\tau_j \leq t \wedge i_j = green\right)}{\sum_{j=1}^N I\left(\tau_j \leq t\right)} \\ yellow_0^t &= \frac{\sum_{j=1}^N I\left(\tau_j \leq t \wedge i_j = yellow\right)}{\sum_{j=1}^N I\left(\tau_j \leq t\right)} \\ radius_0^t &= q_{0.99}\left(\mathcal{A}_0^t\right) \end{aligned} \tag{1}$$

where $green_0^t$ is the fraction of mild seismic reports collected up to time t, $yellow_0^t$ is the fraction of strong seismic reports, $\mathcal{A}_0^t$ is the set of all the spatial distances computed between the reports collected up to time t and $q_{0.99}$ is the $99th$ percentile. The quantity $radius_0^t$ represents a sort of radius of the geographic area impacted by the earthquake. If the earthquake is strong, then $radius_0^t$ is expected to be high and to increase when t increases. The $99th$ percentile is preferred to the maximum in order to mitigate the impact of false reports that, by chance, may be received by the server during a real earthquake. In this context, a report is false when the person sends it to the server without actually feeling an earthquake. It is noted that $1 - green_0^t - yellows_0^t$ is the fraction of very strong seismic reports, which, however, does not add information due to the constraint to one.

There is a loss of information when moving from $\mathcal{D}$ to the quantities in (1). What is lost is the exact time and the exact spatial coordinates of the reports. Nonetheless, (1) are expected to retain useful information for estimating earthquake parameters, while greatly

simplifying the data analysis. For instance, mild seismic reports collected from a large area may be related to a high magnitude but distant earthquake, possibly with epicentre in the sea/ocean and thus far from any populated area. On the other hand, strong seismic reports collected from a small area may be related to a shallow earthquake perceived as strong only near the epicentre. This is the kind of information that (1) are expected to convey.

4. Magnitude model

Before introducing a multivariate model for multiple earthquake parameters, this section details a univariate model for the earthquake magnitude, usually the most important intrinsic parameter for assessing the potential impact of the earthquake on the population.

The aim is modelling the earthquake magnitude at each time $t = 0, ..., T$ given the available information received by the server up to time t. This is done in order to obtain a prediction of the earthquake magnitude and in order to update the prediction while more information is collected by the server.

Three aspects about earthquakes should be discussed. First of all, the magnitude of an earthquake is a measure of the energy released at its hypocentre and an expert validated magnitude is usually available after many minutes from the beginning of the event. On the contrary, people experience the local and instantaneous intensity of the earthquake. The intensity reported by the person's mainly depends on the distance between the person location and the hypocentre of the earthquake.

Second, earthquake hypocentres are usually localised near fault lines. People living in cities close to these faults are expected to experience, on average, strong ground shaking irrespective of the magnitude of the earthquake. On the other hand, cities far from faults will experience mild ground shaking even when earthquakes are strong in magnitude.

Third, people from different countries of the world may rate differently the intensity of an earthquake of a given magnitude. People living in countries characterized by low seismicity usually rate as strong earthquakes which actually have a low magnitude.

The above discussion suggests that any statistical model used for estimating the earthquake magnitude should be characterized, on the one side, by time-variant parameters, and on the other, by a spatial component modelling both the local seismicity and the heterogeneity in the earthquake perception across the world.

The proposed model for the log-magnitude is

$$\begin{aligned} \log y(\boldsymbol{s}, t) &= \boldsymbol{x}_0(t)' \boldsymbol{\beta}_0 + \boldsymbol{x}_1(t)' \boldsymbol{\beta}_1(t) + w(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t) \\ \boldsymbol{\beta}_1(t) &= \boldsymbol{G}\boldsymbol{\beta}_1(t-1) + \boldsymbol{\eta}(t). \end{aligned} \tag{2}$$

In (2), $\boldsymbol{x}_0(t)$ and $\boldsymbol{x}_1(t)$ are vectors of covariates, namely the the quantities (1) detailed in Section 3 and, possibly, any interaction between them. $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1(t)$ are the fixed and time-varying coefficients, respectively, $w(\boldsymbol{s}, t)$ is a latent variable correlated across space and uncorrelated over time while $\varepsilon(\boldsymbol{s}, t) \sim N\left(0, \sigma^2_{\varepsilon,t}\right)$ is a random error with time-varying variance $\sigma^2_{\varepsilon,t}$. Note that the log-transformation is adopted since the histogram of the magnitude of felt earthquakes tends to be right-skewed. Indeed, earthquakes below

magnitude 2.5 are rarely felt and reported by the population. Also, the log-transformation guarantees that a positive magnitude is predicted.

Model (2) is a classic hierarchical model (Gelman and Hill, 2007). In Finazzi et al. (2013), a similar model has been used for modelling the observations of a spatio-temporal phenomenon sampled over time at fixed spatial locations. Here, time t is not the absolute time of a space-time phenomenon but, rather, it represents the evolution over time of the reports related to a generic earthquake and received by the server starting from $t = 0$.

In the simplest case where $\boldsymbol{x}_0(t) = \boldsymbol{x}_1(t)$, $\boldsymbol{\beta}_0$ is the fixed effect for covariates in $\boldsymbol{x}$ while $\boldsymbol{\beta}_1(t)$ is a "correction" of $\boldsymbol{\beta}_0$ specific for time t. Similarly, $w(\boldsymbol{s}, t)$ is a specific effect for spatial location $\boldsymbol{s}$ and time t. Therefore, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1(t)$ and $w(\boldsymbol{s}, t)$ are not related to the physics of the earthquake intended as a spatio-temporal phenomenon, but to the dynamics in space and time of the reports sent by people when an earthquake is felt.

The time-varying coefficients $\boldsymbol{\beta}_1(t)$ have Markovian dynamics with stable diagonal transition matrix $\boldsymbol{G}$ and innovation vector $\boldsymbol{\eta}(t) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$, with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ diagonal. On the other hand, $w(\boldsymbol{s}, t)$ is modelled by means of a Gaussian process with spatial covariance function given by $\rho(\boldsymbol{s}, \boldsymbol{s}'; \theta) = v^2 \exp(-d(\boldsymbol{s}, \boldsymbol{s}')/\theta)$, with v^2 the variance and $d(\cdot)$ the geodetic distance between any two spatial locations $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}^2$.

The model parameter set is $\Psi = \{\boldsymbol{\beta}_0, \boldsymbol{\sigma}_{\varepsilon}^2, v^2, \theta, \boldsymbol{G}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\}$, with $\boldsymbol{\sigma}_{\varepsilon}^2 = \left(\sigma_{\varepsilon,1}^2, ..., \sigma_{\varepsilon,T}^2\right)$.

4.1. Estimate of model parameters

Model estimation is based on the maximum likelihood approach and on the Expectation-Maximization algorithm. In particular, the estimation is based on historical data related to L earthquakes felt and reported by the population through the smartphone app. For each earthquake $l = 1, ..., L$, the following time series

$$\begin{aligned} &\{green_0^t\}_{l,t=1,..,T} \\ &\{yellow_0^t\}_{l,t=1,..,T} \\ &\{radius_0^t\}_{l,t=1,..,T} \end{aligned} \tag{3}$$

are derived from the reports sent to the server. Moreover, $\mathcal{S} = \{\boldsymbol{s}_l\}_{l=1,...,L}$ is the set of all centres of gravity computed at $t = 1$ from the report coordinates received up to $t = 1$.

For each earthquake, its magnitude y_l is obtained from an earthquake catalogue. The data model is then

$$\begin{aligned} \log \boldsymbol{y}_t(\mathcal{S}) &= \boldsymbol{X}_{0,t}\boldsymbol{\beta}_0 + \boldsymbol{X}_{1,t}\boldsymbol{\beta}_{1,t} + \boldsymbol{w}_t(\mathcal{S}) + \boldsymbol{\varepsilon}_t(\mathcal{S}) \\ \boldsymbol{\beta}_{1,t} &= \boldsymbol{G}\boldsymbol{\beta}_{1,t-1} + \boldsymbol{\eta}_t \end{aligned}$$

where $\boldsymbol{y}_t(\mathcal{S}) = (y_1, ..., y_L)'$ is the vector of "true" magnitudes, $\boldsymbol{w}_t(\mathcal{S})$ is the normally distributed spatial latent variable at locations $\mathcal{S}$ and time t, $\boldsymbol{\varepsilon}_t(\mathcal{S})$ is the random error at $\mathcal{S}$ and t while $\boldsymbol{X}_{0,t} = \left(\boldsymbol{x}'_{0,1,t}, ..., \boldsymbol{x}'_{0,L,t}\right)'$ and $\boldsymbol{X}_{1,t} = \left(\boldsymbol{x}'_{1,1,t}, ..., \boldsymbol{x}'_{1,L,t}\right)'$ are the matrices of covariates at time t. The set $\boldsymbol{X} = \{\boldsymbol{X}_{0,1}, ..., \boldsymbol{X}_{0,T}, \boldsymbol{X}_{1,1}, ..., \boldsymbol{X}_{1,T}\}$ collects the covariates at all time steps.

It is worth noting that, contrary to classic spatio-temporal models adopted for modelling evolving phenomena (and thus a time-variant $\boldsymbol{y}_t(\mathcal{S})$), the vector of magnitudes is

time-invariant. This is because the magnitude is a feature of the earthquake assessed a posteriori and that does not change with time. What evolves over time are the model parameters and the latent variables, which are updated at each time t in order to better fit the magnitude on the basis of the available information (covariates) at t.

Details on model estimation using the EM algorithm are given in (Fassò and Finazzi, 2011). Here, it is recalled that the estimate of $\boldsymbol{\beta}_{1,t}$ and its variance are given by $E_{\Psi}\left(\boldsymbol{\beta}_{1,t} \mid \boldsymbol{m}, \boldsymbol{X}\right)$ and $\mathrm{Var}_{\Psi}\left(\boldsymbol{\beta}_{1,t} \mid \boldsymbol{m}, \boldsymbol{X}\right)$, respectively, and they are computed by means of the Kalman smoother (see Shumway and Stoffer (2011)). As common when using the Kalman smoother, $\boldsymbol{\beta}_{1,0} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_0\right)$, with $\boldsymbol{\Sigma}_0$ a diagonal matrix with arbitrarily large variances. Conditional expectation $E_{\Psi}\left(\boldsymbol{w}_t\left(\mathcal{S}\right) \mid \boldsymbol{m}, \boldsymbol{X}\right)$ and the conditional variance $\mathrm{Var}_{\Psi}\left(\boldsymbol{w}_t\left(\mathcal{S}\right) \mid \boldsymbol{m}, \boldsymbol{X}\right)$ are computed using classic formulas of the multivariate normal distribution.

4.2. Prediction of earthquake parameters

The estimated model is eventually used to predict the magnitude of a new earthquake on the basis of the reports collected by the server. For any given spatial location $\boldsymbol{s}^* \in \mathcal{S}^2$ and time step t, the prediction of the log-magnitude is given by

$$\log \hat{y}_t\left(\boldsymbol{s}^*\right) = \boldsymbol{x}'_{0,t}\hat{\boldsymbol{\beta}}_0 + \boldsymbol{x}'_{1,t}\hat{\boldsymbol{\beta}}_{1,t} + \hat{w}_t\left(\boldsymbol{s}^*\right) \tag{4}$$

where $\boldsymbol{s}^*$ is the centre of gravity of the reports collected at $t = 1$ while $\hat{w}_t\left(\boldsymbol{s}^*\right)$ is the latent variable $w\left(\boldsymbol{s}, t\right)$ estimated at spatial location $\boldsymbol{s}^*$ by means of classic spatial Kriging (see Diggle et al. (1998)). The log-magnitude estimation variance is given by

$$\begin{aligned}\hat{\sigma}^2_y\left(\boldsymbol{s}^*\right) = \underset{\hat{\Psi}}{\mathrm{Var}}\left(y_t\left(\boldsymbol{s}^*\right) \mid \boldsymbol{y}, \boldsymbol{X}\right) = \boldsymbol{x}'_{1,t}\underset{\hat{\Psi}}{\mathrm{Var}}\left(\boldsymbol{\beta}_{1,t} \mid \boldsymbol{y}, \boldsymbol{X}\right)\boldsymbol{x}_{1,t} + \underset{\hat{\Psi}}{\mathrm{Var}}\left(w_t\left(\boldsymbol{s}^*\right) \mid \boldsymbol{y}, \boldsymbol{X}\right) \\ + 2\boldsymbol{x}'_{1,t}cov_{\hat{\Psi}}\left(\boldsymbol{\beta}_{1,t}, w_t\left(\boldsymbol{s}^*\right) \mid \boldsymbol{y}, \boldsymbol{X}\right).\end{aligned} \tag{5}$$

In (5), conditional variance and covariance are based on the estimated parameters $\hat{\Psi}$. Also note that $\boldsymbol{x}_{0,t}$ and $\boldsymbol{x}_{1,t}$ are the vectors of covariates for the current earthquake while $\boldsymbol{X}$ refers to the historical data. Thanks to the estimation variance, 95% confidence intervals for the log-magnitude are easily obtained as $\hat{y}_t\left(\boldsymbol{s}^*\right) \pm 1.96\sqrt{\hat{\sigma}^2_y\left(\boldsymbol{s}^*\right)}$.

Using (4) and (5), it is possible to update the magnitude prediction and its confidence interval while new reports are sent to the server. The update is done with temporal interval Δ.

5. Multivariate model

The model described in the previous section is easily extended to more than one earthquake parameter. Let $\boldsymbol{y}$ be the $k \times 1$ vector of earthquake parameters of interest. The multivariate model for $\log \boldsymbol{y}$ is

$$\begin{aligned}\log \boldsymbol{y}\left(\boldsymbol{s}, t\right) &= \boldsymbol{X}_0\left(t\right)\boldsymbol{\beta}_0 + \boldsymbol{X}_1\left(t\right)\boldsymbol{\beta}_1\left(t\right) + \boldsymbol{w}\left(\boldsymbol{s}, t\right) + \boldsymbol{\varepsilon}\left(\boldsymbol{s}, t\right) \\ \boldsymbol{\beta}_1\left(t\right) &= \boldsymbol{G}\boldsymbol{\beta}_1\left(t-1\right) + \boldsymbol{\eta}\left(t\right)\end{aligned} \tag{6}$$

where $\boldsymbol{X}_r(t) = I_k \otimes \boldsymbol{x}_r(t)'$ is the $k \times kb$ block-diagonal matrix of covariates, $r = 0, 1$, $\boldsymbol{w}(\boldsymbol{s}, t)$ is a multivariate Gaussian process with matrix spatial covariance function given by $\rho(\boldsymbol{s}, \boldsymbol{s}'; \theta) = \boldsymbol{V} \exp(-d(\boldsymbol{s}, \boldsymbol{s}')/\theta)$ while $\boldsymbol{\varepsilon}(\boldsymbol{s}, t) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\varepsilon,t})$ is the random error. Matrix $\boldsymbol{V}$ is a valid $k \times k$ variance-covariance matrix while $\boldsymbol{\Sigma}_{\varepsilon,t}$ is a $k \times k$ time-varying diagonal matrix with diagonal elements collected in vector $\boldsymbol{\sigma}^2_{\varepsilon}$. Vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1(t)$ have kb elements while $\boldsymbol{G}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ are diagonal $kb \times kb$ matrices. In (6), it is assumed that the same vector of covariates $\boldsymbol{x}(t)$ is used to model each earthquake parameter. In practice, each earthquake parameter can have specific covariates. The model parameter set becomes $\Psi = \{\boldsymbol{\beta}_0, \boldsymbol{\sigma}^2_{\varepsilon}, \boldsymbol{V}, \theta, \boldsymbol{G}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\}$. Model estimation and prediction of the earthquake parameters are analogous to those detailed in the previous section. Details on the estimate of the multivariate model are given in Fassò and Finazzi (2011).

6. Case study

In this section, a case study is developed in order to show the fitting capabilities of model (6) and to show how the model is used for real time prediction and updating. The focus of the case study is on the estimate of three parameters: two intrinsic earthquake parameters, namely magnitude and depth, and the distance between the earthquake epicentre and the centre of gravity of the report coordinates at $t = 1$, here called d_1. Being a distance, d_1 is not a prediction of the epicentre location but it is useful to assess the impact of the earthquake on the population. As an example, a magnitude 8 earthquake with epicentre at 10 km from the area of the first reports is expected to have a higher impact than an earthquake of the same magnitude but 500 km away. A statistical model for the real time estimate of the epicentre from smartphone network data is detailed in Finazzi (2016). However, temporal and spatial information of the reports must be retained in order to obtain unbiased predictions, especially if the epicentre is "outside" the network of smartphones. For this reason, epicentre estimation is outside the scope of this work and distance d_1 is used as a proximity measure of the earthquake to the nearest populated area.

6.1. Data description

The data set considered for the case study consists in $222,288$ reports collected through the Earthquake Network app in the period April 20, 2014 - November 17, 2018. Reports relate to $1,552$ earthquakes identified by matching the reports to the earthquakes listed in the United States Geological Survey (USGS) catalogue‡.

Figure 2 shows the spatial distribution of all the reports collected in this period. The distribution highlights some of the major seismic zones of the world, with most of the reports coming from Central and South America but also Italy, Nepal, Taiwan, Philippines, Indonesia and California in United States of America.

Figure 3 depicts the $1,552$ earthquakes in terms of their magnitude and the distance d_1. As expected, the distance tends to be small for earthquakes of low magnitude.

On the other hand, Figure 4 shows the earthquake magnitude versus the number of reports sent to the server. Note that a high magnitude does not necessarily imply a large

‡The catalogue is available online at https://earthquake.usgs.gov/earthquakes/search/

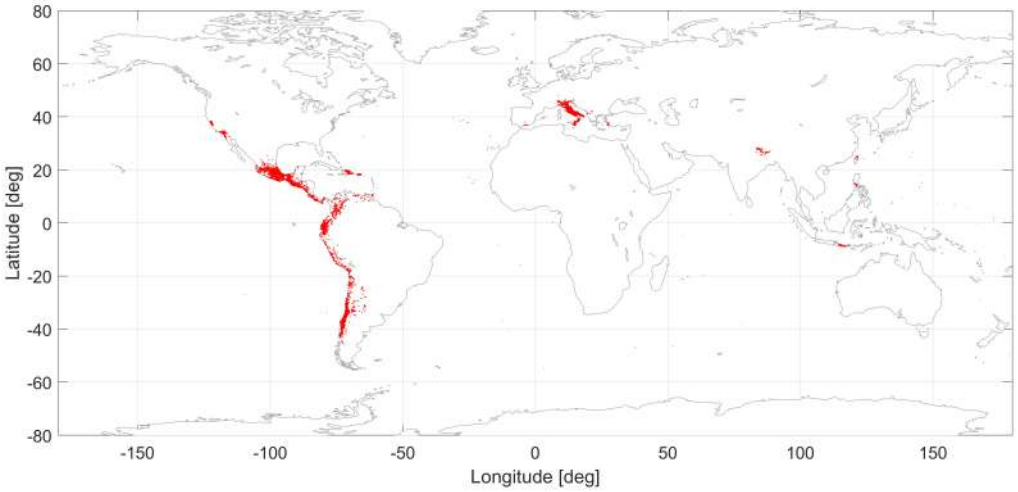

Fig. 2. Spatial distribution of the $222,288$ seismic reports sent to the Earthquake Network server in the period April 20, 2014 - November 17, 2018. Each dot is a report.

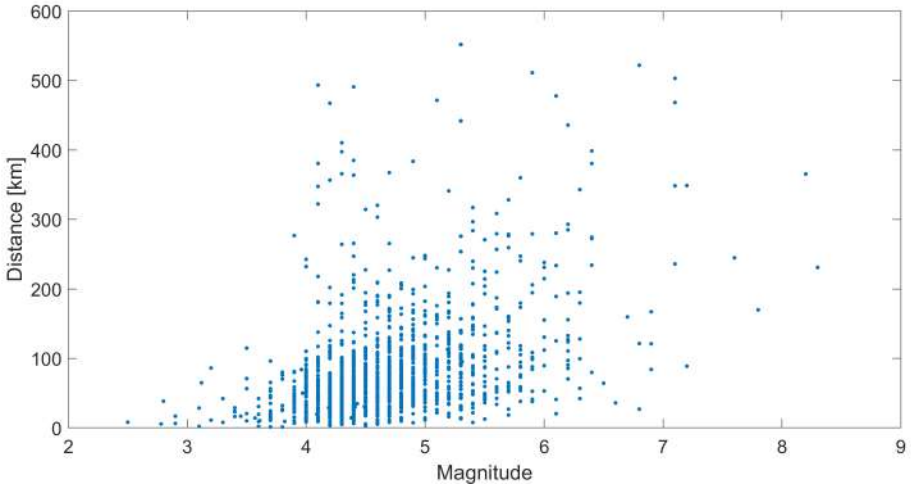

Fig. 3. Earthquake magnitude vs distance d_1 for the $1,552$ earthquakes of the case study.

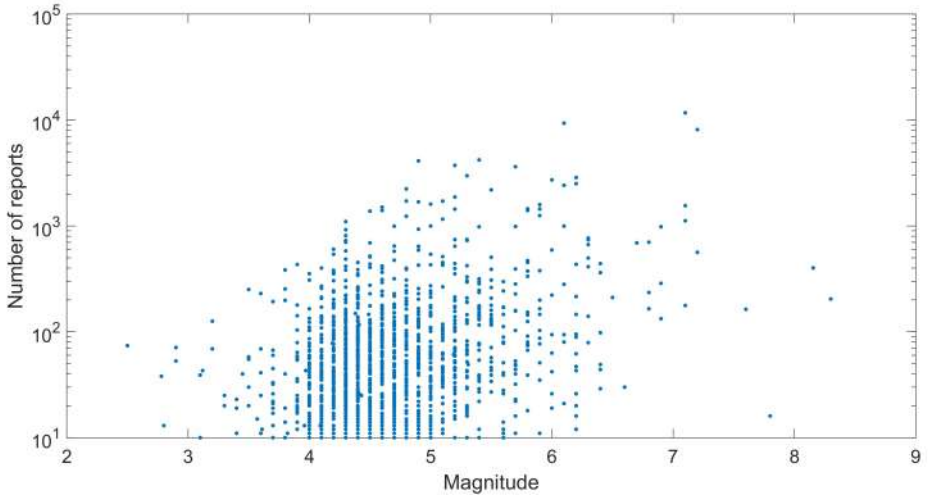

Fig. 4. Earthquake magnitude vs number of seismic reports for the $1,552$ earthquakes of the case study.

number of reports. In fact, the number of reports depends on the distance from the epicentre and on the spread of the app among the population.

For this work, it is decided to study the evolution of the reports over a period of 10 minutes considering $\Delta = 10$ seconds as time interval. This implies $t = 0, ..., 60$. For each earthquake, then, time series (3) are computed from the data set $\mathcal{D}$. Figure 5 shows the values of $green_0^t$, $yellow_0^t$ and $radius_0^t$ versus the earthquake magnitude when $t = 20$. Analogous figures for depth and distance d_1 are given in supplementary material.

The choice $t = 20$ reflects the fact that, after 200 seconds, the reports sent to the server should well represent the earthquake perception by the population.

Table 2 reports the linear correlations between $green_0^t$, $yellow_0^t$ and $radius_0^t$ and the three earthquake parameters. The ratios of mild and strong reports show a correlation with magnitude but not with depth and distance, while the radius presents a correlation with all the parameters. As expected, the linear correlation between magnitude and $green_0^t$ is negative, since the number of mild seismic reports is expected to be low when the earthquake is strong. On the contrary, the linear correlations between magnitude and both $yellow_0^t$ and $radius_0^t$ are positive since a strong earthquake implies a higher number of strong seismic reports and a larger radius for the reports. Although small, the linear correlation between depth and $radius_0^t$ is positive. If the earthquake is deep, it must be strong to be felt at all by the population, and being strong is felt over a large area. Instead, there is no easy or obvious explanation for the positive linear correlation between $radius_0^t$ and distance, which however is small.

6.2. Model estimation and comparison

Since three parameters are of interest, the subsequent data analysis will be based on the multivariate model (6). All the parameters are log-transformed. Additionally, variable

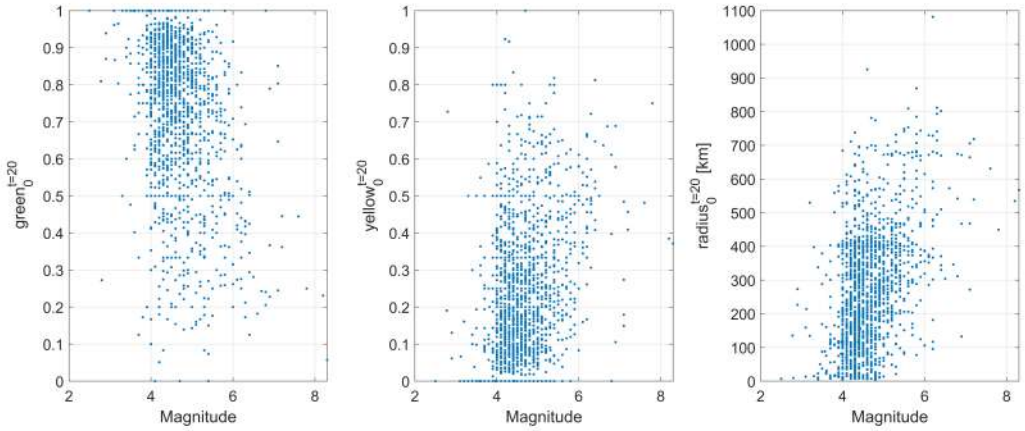

Fig. 5. Values of $green_0^t$, $yellow_0^t$ and $radius_0^t$ vs earthquake magnitude when $t = 20$ for the $1,552$ earthquakes of the case study.

Table 2. Linear correlations between quantities $green_0^t$, $yellow_0^t$, $radius_0^t$ and magnitude, depth and distance d_1 when $t = 20$.

	Magnitude	*Depth*	*Distance*
$green_0^t$	-0.34	0.04	0.02
$yellow_0^t$	0.30	-0.03	-0.08
$radius_0^t$	0.44	0.19	0.23

Table 3. Estimated fixed effects $\hat{\beta}_0$ for the multivariate model (6). Standard deviations in brackets.

	Magnitude	*Depth*	*Distance*
green	$-0.156_{(0.155)}$	$0.096_{(0.012)}$	$-1.078_{(0.061)}$
yellow	$1.298_{(0.081)}$		$-1.958_{(0.060)}$
radius	$0.397_{(0.016)}$		$0.255_{(0.005)}$
green:yellow	$-0.259_{(0.023)}$		$0.598_{(0.023)}$
green^2	$0.330_{(0.050)}$	$-0.032_{(0.011)}$	$0.538_{(0.050)}$
yellow^2	$-0.746_{(0.030)}$		$0.987_{(0.030)}$

and covariates are normalised to have zero mean and unit variance. This helps numerical stability during model estimation and it makes the $\boldsymbol{\beta}$ coefficients directly comparable.

To show the benefit of using a complex model with temporal and spatial latent variables, two additional simpler models are considered. The first is a classic regression model without time-varying coefficients and without the spatial latent variable:

$$\log y_i(\boldsymbol{s}, t) = \boldsymbol{x}_0(t)' \boldsymbol{\beta}_0 + \varepsilon(\boldsymbol{s}, t) \tag{7}$$

while the second is a dynamic linear model with only time-varying coefficients:

$$\log y_i(\boldsymbol{s}, t) = \boldsymbol{x}_0(t)' \boldsymbol{\beta}_0 + \boldsymbol{x}_1(t)' \boldsymbol{\beta}_1(t) + \varepsilon(\boldsymbol{s}, t) \tag{8}$$

where $\boldsymbol{\beta}_1(t)$ has the same Markovian dynamics in (2) and where the index i runs over the three earthquake parameters. For brevity, $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ will refer to (7), (8) and (6), respectively.

Model predictive capabilities are assessed using cross-validation. In particular, 75% of the $1,552$ earthquakes is randomly sampled and used for model estimation while the remaining 25% is used for cross-validation.

For each variable, a preliminary analysis not reported here has been carried out in order to identify the significant covariates (and interactions) to be included in $\boldsymbol{x}_0(t)$ and $\boldsymbol{x}_1(t)$. Model estimation is carried out using the D-STEM software (Finazzi and Fassò, 2014). D-STEM provides standard deviations for the estimated model parameters enabling inference.

The estimated parameters are reported in Tables 3-4 and in Figure 6, and they already reflect the pruning of the preliminary analysis. Table 4 reports the estimated variance-covariance matrix $\hat{\boldsymbol{V}}$ of the multivariate spatial latent variable $\boldsymbol{w}(\boldsymbol{s}, t)$. Covariance is high between magnitude and distance. This is likely due to the fact that high magnitude earthquakes are felt at large distance. The estimated parameter of the spatial correlation function ρ is $\hat{\theta} = 107.6\ km$ with standard deviation $0.1\ km$. This implies that $\boldsymbol{w}(\boldsymbol{s}, t)$ is essentially uncorrelated for any two points more than $300\ km$ apart but, for instance, is highly correlated within the same city or large metropolitan area.

Figures 6-8 show the estimated time-varying effects $\hat{\beta}_0 + \hat{\beta}_1(\Delta \cdot t)$ (only those significantly different from zero) for covariates used to explain log-magnitude, log-depth and log-distance, respectively. Graphs show how the effect of each covariate changes with respect to the time elapsed since the first report (i.e., $t = 0$). This reflects the fact that the information content of the reports changes with time and that the model has to adapt its parameters.

Table 4. Estimated variance-covariance matrix $\hat{V}$ of the latent spatial variable $\boldsymbol{w}(\boldsymbol{s}, t)$ for the multivariate model (6). Standard deviations in brackets.

	Magnitude	*Depth*	*Distance*
Magnitude	$0.59_{(5.8 \cdot 10^{-4})}$	$0.02_{(3.3 \cdot 10^{-4})}$	$0.55_{(4.2 \cdot 10^{-4})}$
Depth	$0.02_{(3.3 \cdot 10^{-4})}$	$0.58_{(4.3 \cdot 10^{-4})}$	$0.15_{(3.4 \cdot 10^{-4})}$
Distance	$0.55_{(4.2 \cdot 10^{-4})}$	$0.15_{(3.4 \cdot 10^{-4})}$	$0.72_{(6.0 \cdot 10^{-4})}$

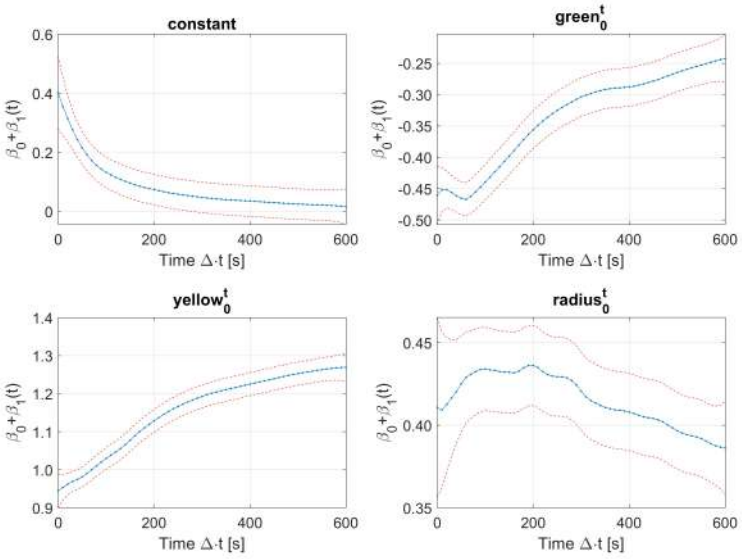

Fig. 6. Time-varying effect $\hat{\beta}_0 + \hat{\beta}_1 (\Delta \cdot t)$ and 95% confidence interval (dotted lines) for covariates used to explain the log-magnitude.

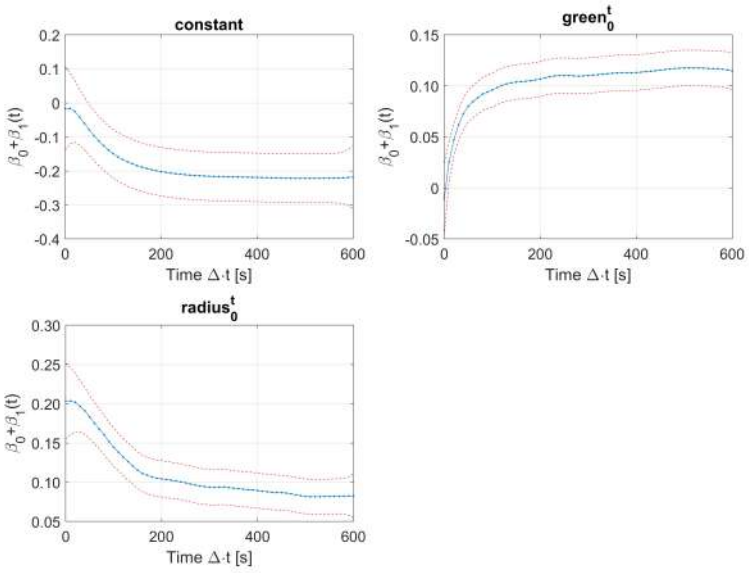

Fig. 7. Time-varying effect $\hat{\beta}_0 + \hat{\beta}_1 (\Delta \cdot t)$ and 95% confidence interval (dotted lines) for covariates used to explain the log-depth.

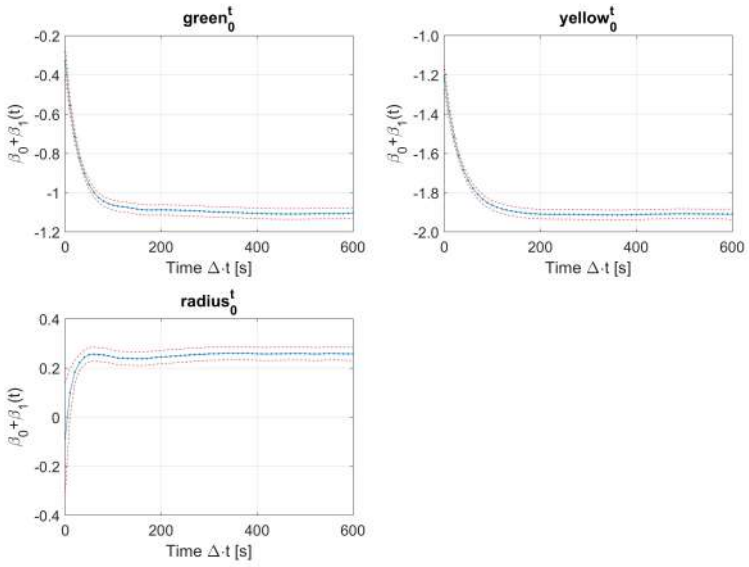

Fig. 8. Time-varying effect $\hat{\beta}_0 + \hat{\beta}_1 (\Delta \cdot t)$ and 95% confidence interval (dotted lines) for covariates used to explain the log-distance.

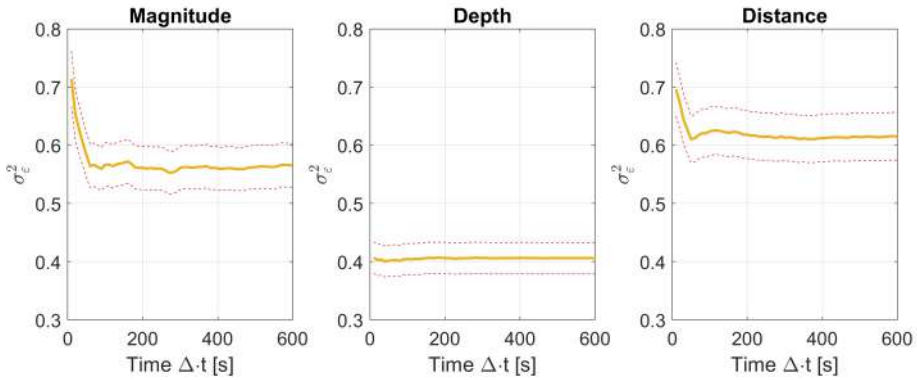

Fig. 9. Estimated $\hat{\sigma}^2_{\varepsilon,t}$ and 95% confidence interval (dotted lines) of multivariate model (6) for magnitude, depth and distance d_1.

Finally, Figure 9 shows the estimated $\hat{\sigma}^2_{\varepsilon,t}$ and 95% interval for the three earthquake parameters. For magnitude and distance, the variance of the random error $\varepsilon(\boldsymbol{s}, t)$ is significantly higher for the first time steps while it stabilises as t increases. For depth, the variance is essentially constant over time.

Figure 10 shows $\hat{w}_t(\boldsymbol{s}^*)$ and its variance $\mathrm{Var}_{\hat{\Psi}}(w_t(\boldsymbol{s}^*) \mid \boldsymbol{y}, \boldsymbol{X})$ for magnitude and for two selected regions of the world. In this case, $\boldsymbol{s}^*$ belongs to a regular spatial grid of 0.5° resolution. Recall that w is a latent variable describing a spatial local effect not explained by covariates. A positive (negative) w means that covariates tend to underpredict (overpredict) the magnitude. As an example, the top-left panel of Figure 10 depicts $\hat{w}_t(\boldsymbol{s}^*)$ over a sub-region of South America at $t = 60$. Note that $\hat{w}$ is mainly negative in Chile while it is positive in the region of the city of Cordoba in Argentina. This can be explained by the fact that most of the earthquakes (including very strong earthquakes) occur in the Pacific Ocean along the coast of Chile. Strong earthquakes in Chile are also felt in Argentina but they are reported as mild due to the large distance from the epicentre. Being positive, $\hat{w}$ allows to compensate for this and to predict a higher magnitude. The bottom-left panel depicts $\hat{w}$ for Italy. During the period covered by this work, four earthquakes with magnitude between 5.4 and 6.5 hit Central Italy. In particular, the magnitude 6.0 earthquake that hit on August 24, 2016 killed 299 people. All these earthquakes were reported as very strong by the users Earthquake Network app, even though their magnitudes were not high if compared to global seismicity. In this case, $\hat{w}$ suggests to decrease the prediction which is made using only covariates. This, in turn, shows that $\hat{w}$ also describes a global heterogeneity in the perception of the earthquake intensity by the population. Right panels in Figure 10 represent the variance of $\hat{w}$, namely its uncertainty. Uncertainty is low (high) in regions where a large (small) number of reports where observed and used for model estimation. Moreover, the higher the uncertainty on $\hat{w}$ the wider the confidence interval on the magnitude prediction. Analogous figures for the entire globe and for the three earthquake parameters are given as supplementary material.

Figures 11 and 12 depict back-transformed $\hat{\boldsymbol{y}}$ versus $\boldsymbol{y}$ when $t = 60$ for data used to fit the model (in sample data) and for cross-validation data, respectively. For all earthquake

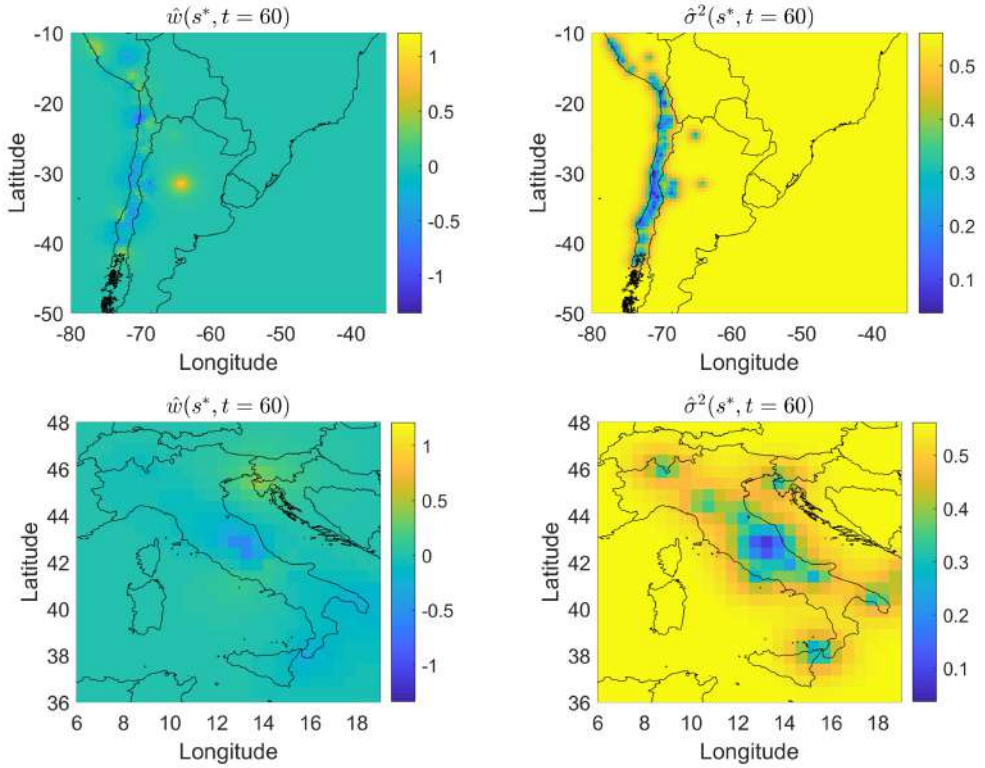

Fig. 10. Spatial latent variable $\hat{w}_t(\boldsymbol{s}^*)$ (left panels) and its variance $\mathrm{Var}_{\hat{\Psi}}(w_t(\boldsymbol{s}^*) \mid \boldsymbol{m}, \boldsymbol{X})$ (right panels) estimated for a sub-region of South America (top panels) and Italy (bottom panels) when $t = 60$.

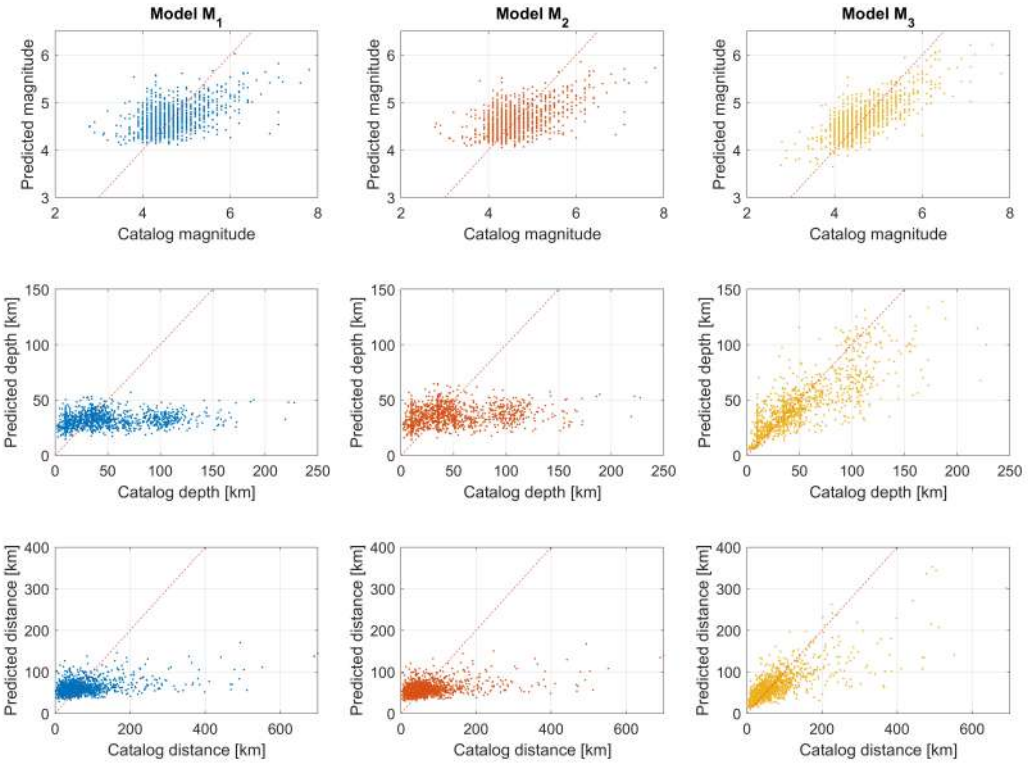

Fig. 11. Scatter plots of back-transformed $\hat{\boldsymbol{y}}$ versus $\boldsymbol{y}$ when $t = 60$ for in sample data.

parameters, the improvement in the model fitting capability is visible when moving from $\mathcal{M}_1$ to $\mathcal{M}_3$. Table 5 gives the R^2 statistics when $t = 60$ for all models and all parameters. Again, the benefit of using model $\mathcal{M}_3$ is clear. Looking at Figures 11, it is possible to see that there is a tendency to overpredict (underpredict) the magnitude of low (high) magnitude earthquakes. This is possibly due to the fact that the dataset includes a small number of both low and high magnitude earthquakes, and model parameters are thus optimized for medium magnitude earthquakes. In fact, high magnitude earthquakes are rare while small magnitude earthquakes are large in number but they are rarely felt and/or reported by the population.

Figure 13 shows the trend of the in sample and of the cross-validation root mean squared error (RMSE) for all the parameters (back-transformed) and all the models. Besides the improvement related to model $\mathcal{M}_3$, it is possible to note how the RMSE decreases over time. For instance, the magnitude RMSE stabilizes after 50 seconds from the first report received by the server. A similar behaviour is observed for depth and distance. In cross-validation, the magnitude RMSE reaches its minimum at around 170 seconds and then increases. This is possibly due to the fact that, after few minutes from the beginning of the earthquake, the server starts receiving false reports, usually from people who are alerted by the smartphone app and who send a report without

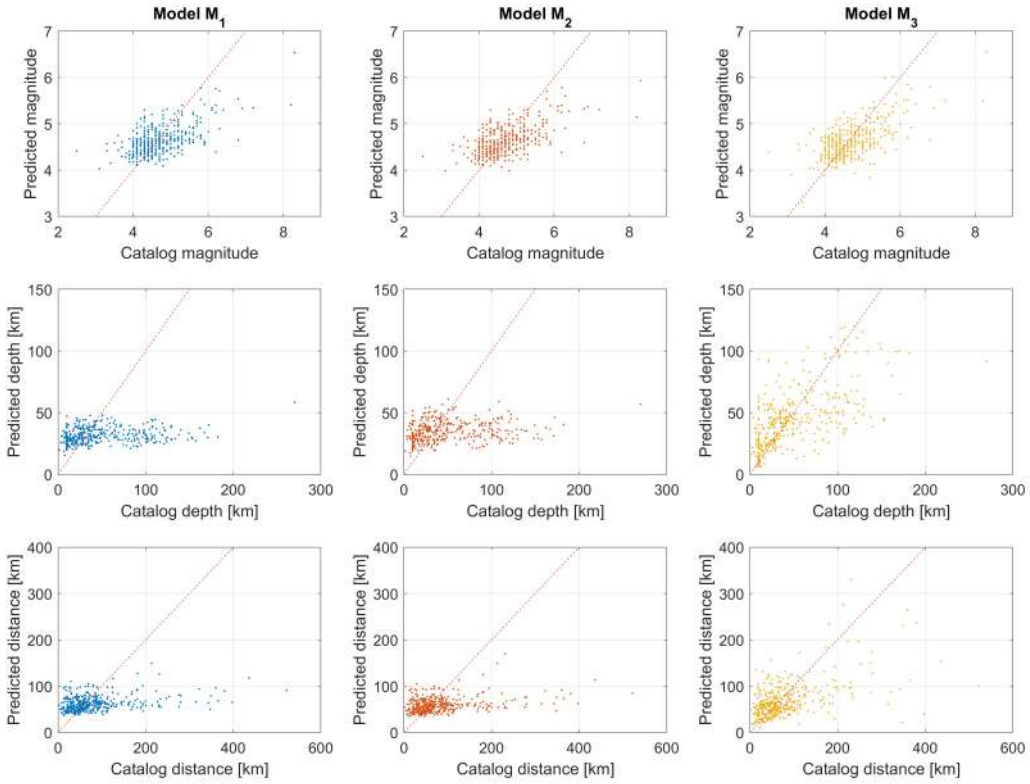

Fig. 12. Scatter plots of back-transformed $\hat{\boldsymbol{y}}$ versus $\boldsymbol{y}$ when $t = 60$ for cross-validation data.

actually experiencing the earthquake. This suggests that after 170 seconds the magnitude prediction is less reliable.

6.3. Prediction of the earthquake parameters

Once estimated, the model is used to analyse the seismic reports collected during new events. Earthquake parameters are predicted and updated with temporal interval Δ while reports keep arriving at the server. As an example, Figure 14 shows the prediction over time for the magnitude 5.2 earthquake the reports of which are shown in Figure 2. Graphs in Figure 14 show how the earthquake parameters are updated while time passes, confidence intervals included. Note that the prediction variance (5) also depends on the variance of the estimated spatial variable $\hat{w}(\boldsymbol{s}, t)$. This variance is higher for areas of the world where few or no reports where observed; therefore, confidence intervals are also expected to be wider for those areas.

Finally, Figure 15 depicts on a map the evolution over time of the predicted distance d_1 and its 95% confidence interval. At $t = 1$, the seismic reports collected within the first $\Delta = 10$ seconds are used to compute the centre of gravity $\boldsymbol{s}$ of their coordinates (asterisk marker in the figure panels). While new reports are collected, the prediction of d_1 and its

Table 5. In sample and cross-validation R^2 for magnitude, depth and distance d_1 under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ at $t = 60$.

	In sample			*Cross-validation*		
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
Magnitude	0.26	0.27	0.51	0.35	0.33	0.42
Depth	0.02	0.02	0.62	0.02	0.02	0.40
Distance	0.10	0.10	0.47	0.07	0.08	0.27

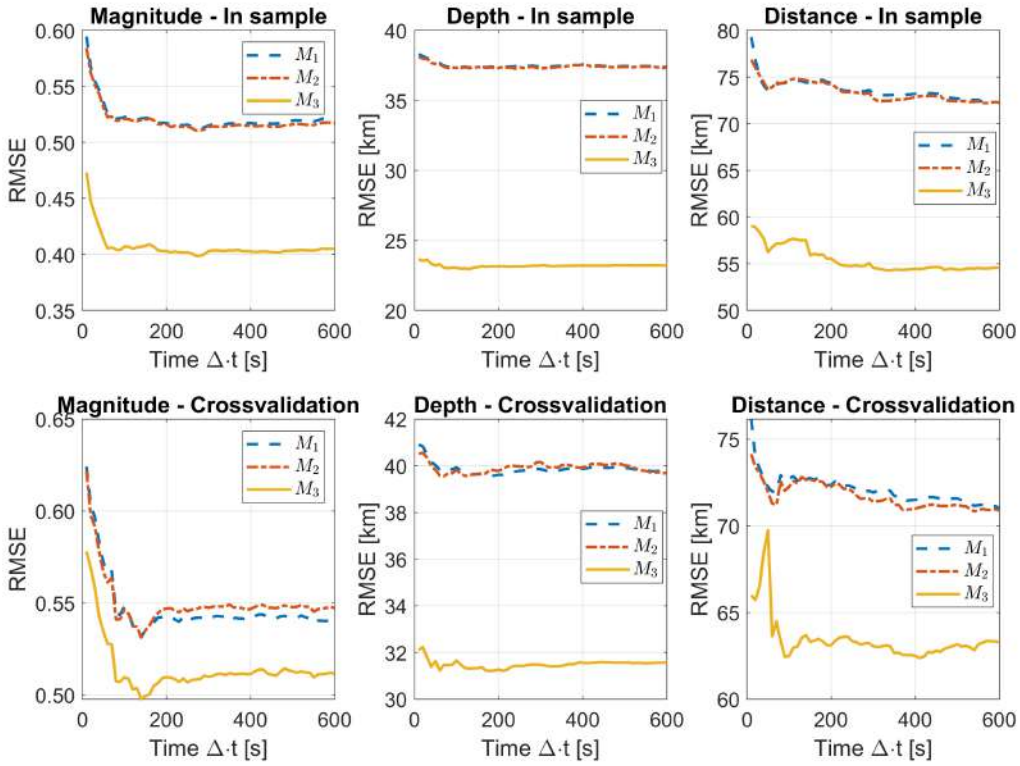

Fig. 13. In sample and cross-validation RMSE for magnitude, depth and distance d_1 under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$.

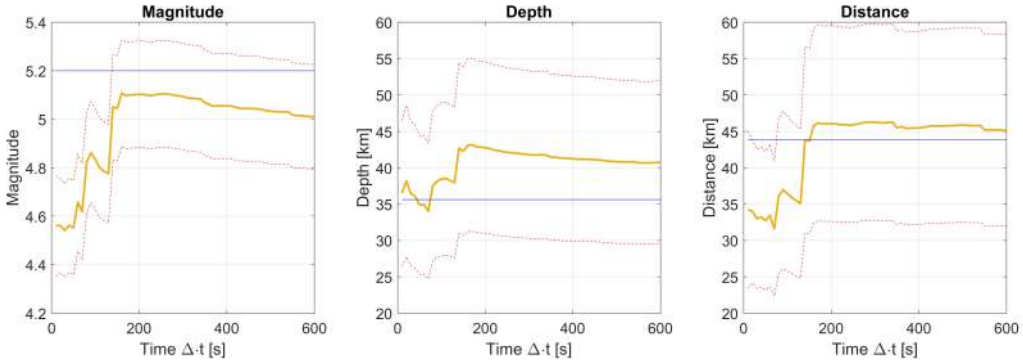

Fig. 14. Model prediction for the magnitude 5.2 earthquake detailed in Figure 2. Thick line: estimation; light line: real earthquake parameter from USGS earthquake catalogue; dotted line: 95% confidence interval.

confidence interval are updated. Note that only d_1 is updated while the centre of gravity $\boldsymbol{s}$ is fixed once computed at $t = 1$. Moreover, it is possible to see that the first prediction for d_1 is already a good estimate of the distance between $\boldsymbol{s}$ and the earthquake epicentre. This is possible thanks to the spatial variable w which, in this case, well describes the typical distance of earthquake epicentres when reports has centre of gravity at $\boldsymbol{s}$.

7. Remarks

Results obtained from the data analysis of the previous section allow us to answer the four questions posed in Section 1.

(a) Can seismic reports be used to predict intrinsic parameters of an earthquake in real time? Results suggest that a statistical model with time-varying coefficients and a spatial latent variable can be fitted in order to provide good prediction of earthquake parameters useful to assess the impact on the population.
(b) How reliable are such predictions? For magnitude, depth and distance, the in sample RMSE is around 0.4, 23 km and 55 km, respectively. The cross-validation RMSE is around 20%, 35% and 15% higher, respectively.
(c) After how many seconds are the predictions reliable? Predictions become reliable and stable after around 50 seconds for magnitude and depth and after around 200 seconds for distance.
(d) Do predictions keep improving with time? Once stable, there is no significant improvement in the predictions. Predictions for magnitude possibly deteriorate after 170 seconds.

8. Conclusions

In this paper, seismic intensity reports collected by the Earthquake Network citizen science project have been analysed in order to understand if the report information content

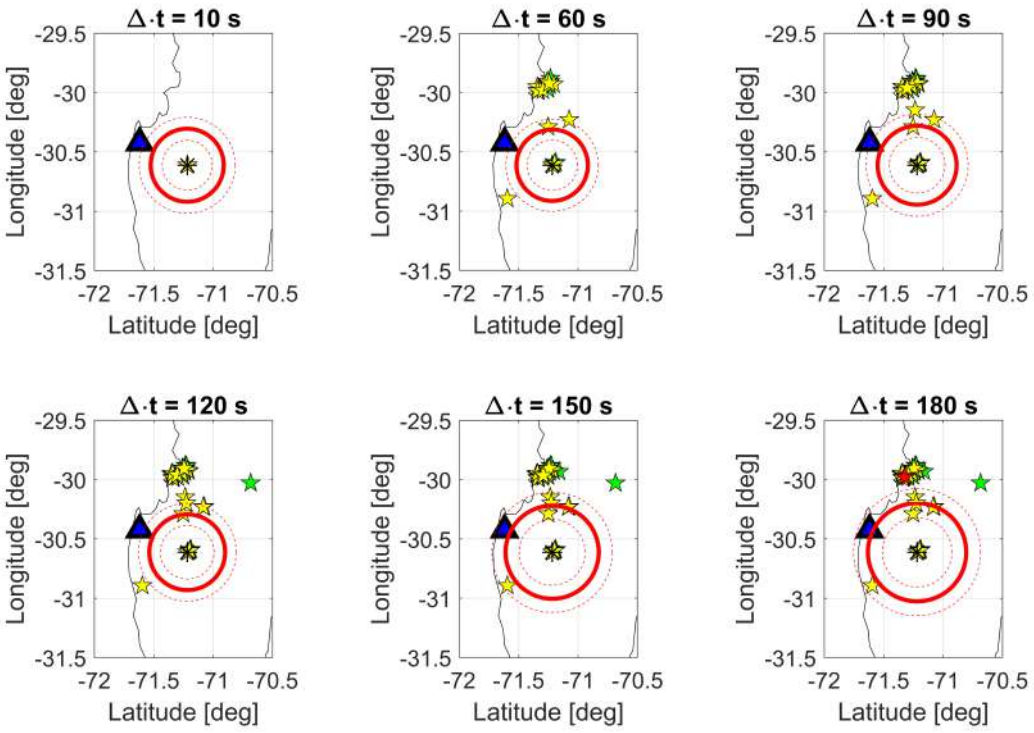

Fig. 15. Prediction of distance d_1 for the magnitude 5.2 earthquake detailed in Figure 2. The triangular marker is the earthquake epicentre while star markers are the locations of the seismic reports. The asterisk marker is the centre of gravity $\boldsymbol{s}$ of the coordinates of the seismic reports collected within the first $\Delta = 10$ seconds. The thick line circle is centred on $\boldsymbol{s}$ and has radius d_1, while the dotted lines represent the 95% confidence interval on d_1.

is suitable for predicting earthquake parameters while the earthquake is evolving. Predictions are eventually communicated in near real time to the general public through smartphone apps, fulfilling the need for information arising among the population.

Parameters of interest were magnitude, depth and distance between the earthquake epicentre and the location of the first reports sent to the server. The modelling of the parameters was done using a multivariate space-time model with time-varying coefficients and a latent spatial variable. The time-varying coefficients allows the model parameters to be adapted to a report information content that changes with time, while the latent spatial variable is able to capture the local seismicity and the global heterogeneity in the perception of the earthquake intensity by the population.

The model has been applied to more than 200 thousand seismic reports globally collected over a period of more than 4 years. For all the earthquake parameters, it was shown that the model proposed in the paper outperforms a classic regression model and a model with time-varying coefficients but without the spatial variable. Thanks to the model, it was possible to assess prediction root mean squared error and the amount of time that has to be waited before obtaining reliable predictions. Moreover, it was shown that waiting longer does not significantly improve predictions, both when considering in sample data and when considering cross-validation data.

Cross-validation RMSE are neither too large for the predictions to be uninformative nor small enough to claim that they can replace the official estimates coming from seismic networks. Nonetheless, predictions provided by the approach developed in this paper have the nature of preliminary information and they have the advantage of being in near real time. Moreover, it has been shown that the need for information can be fulfilled by the population itself by providing information about what they are experiencing.

Future works will consider seismic reports coming from multiple citizen science projects, which will be jointly analysed to improve the real time prediction, solving the problem of possibly misaligned seismic intensity scales. Moreover, exploiting insights from other citizen science projects discussed in Bain (2016), the possibility of assigning a personal measure of reliability to people joining the projects will be exploited. Such a measure will allow to weight differently the information coming from different smartphone users, with a higher weight assigned to trusted users and/or users with a long history in the project.

References

Allen, R. M. and H. Kanamori (2003). The potential for earthquake early warning in Southern California. *Science 300*(5620), 786–789.

Allen, T. I., D. J. Wald, P. S. Earle, K. D. Marano, A. J. Hotovec, K. Lin, and M. G. Hearne (2009). An atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bulletin of Earthquake Engineering 7*(3), 701–718.

Atkinson, G. M. and D. J. Wald (2007). "Did You Feel It?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismological Research Letters 78*(3), 362–368.

Bain, R. (2016). Citizen science and statistics: Playing a part. *Significance 13*(1), 16–21.

Bossu, R., M. Laurin, G. Mazet-Roux, F. Roussel, and R. Steed (2015). The importance of smartphones as public earthquake-information tools and tools for the rapid engagement with eyewitnesses: A case study of the 2015 Nepal earthquake sequence. *Seismological Research Letters 86*(6), 1587.

Bossu, R., F. Roussel, L. Fallou, M. Landès, R. Steed, G. Mazet-Roux, A. Dupont, L. Frobert, and L. Petersen (2018). LastQuake: From rapid information to global seismic risk reduction. *International Journal of Disaster Risk Reduction 28*, 32 – 42.

Cameletti, M., V. De Rubeis, C. Ferrari, P. Sbarra, and P. Tosi (2017, Sep). An ordered probit model for seismic intensity data. *Stochastic Environmental Research and Risk Assessment 31*(7), 1593–1602.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS 17*(1), 124–147.

De Rubeis, V., C. Gasparini, and P. Tosi (1992). Determination of the macroseismic field by means of trend and multivariate analysis of questionnaire data. *Bulletin of the Seismological Society of America 82*(3), 1206–1222.

Diggle, P. J., J. Tawn, and R. Moyeed (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*(3), 299–350.

Earle, P. S., D. C. Bowden, and M. Guy (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics 54*(6).

Fassò, A. and F. Finazzi (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics 22*(6), 735–748.

Finazzi, F. (2016). The earthquake network project: Toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America 106*(3), 1088–1099.

Finazzi, F. and A. Fassò (2014). D-STEM: a software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software, Articles 62*(6), 1–29.

Finazzi, F. and A. Fassò (2017, Sep). A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stochastic Environmental Research and Risk Assessment 31*(7), 1649–1658.

Finazzi, F., E. M. Scott, and A. Fassò (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(2), 287–308.

Gehl, P., F. Cavalieri, and P. Franchin (2018). Approximate bayesian network formulation for the rapid loss assessment of real-world infrastructure systems. *Reliability Engineering & System Safety 177*, 80 – 93.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Graham, M., S. A. Hale, and D. Gaffney (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer 66*(4), 568–578.

Jordan, T., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011). Operational earthquake forecasting. State of knowledge and guidelines for utilization. *Annals of Geophysics 54*(4).

Sakaki, T., M. Okazaki, and Y. Matsuo (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM.

Sbarra, P., P. Tosi, and V. De Rubeis (2010, Aug). Web-based macroseismic survey in Italy: method validation and results. *Natural Hazards 54*(2), 563–581.

Shumway, R. H. and D. S. Stoffer (2011). Time series regression and exploratory data analysis. In *Time series analysis and its applications*, pp. 47–82. Springer.

Silva, V., D. Amo-Oduro, A. Calderon, J. Dabbeek, V. Despotaki, L. Martins, A. Rao, M. Simionato, D. Viganò, C. Yepes, A. Acevedo, N. Horspool, H. Crowley, K. Jaiswal, M. Journeay, and M. Pittore (2018). Global Earthquake Model (GEM) exposure map (version 2018.1).

Tosi, P., P. Sbarra, V. De Rubeis, and C. Ferrari (2015). Macroseismic intensity assessment method for web questionnaires. *Seismological Research Letters 86*(3), 985–990.

Wald, D. J., V. Quitoriano, T. H. Heaton, H. Kanamori, C. W. Scrivner, and C. B. Worden (1999). Trinet "ShakeMaps": Rapid generation of peak ground motion and intensity maps for earthquakes in Southern California. *Earthquake Spectra 15*(3), 537–555.