Random Forest location prediction from social networks during disaster events.

Rachid Ouaret, B. Birregah, Eddie Soulier, Samuel Auclair, Faïza Boulahya

Random Forest location prediction from social networks during disaster events

Rachid Ouaret
Charles Delaunay Institute, UMR CNRS 6281
University of Technology of Troyes
Troyes, France
rachid.ouaret@utt.fr

Babiga Birregah
Charles Delaunay Institute, UMR CNRS 6281
University of Technology of Troyes
Troyes, France
babiga.birregah@utt.fr

Eddie Soulier
Tech-CICO Team
University of Technology of Troyes
Troyes, France
eddie.soulier@utt.fr

Samuel Auclair
French Geological Survey
BRGM
Orléans, France
s.auclair@brgm.fr

Faiza Boulahya
French Geological Survey
BRGM
Orléans, France
F.Boulahya@brgm.fr

***Abstract*—Rapid location and classification of data posted on social networks during time-critical situations such as natural disasters, crowd movement and terrorism is very useful way to gain situational awareness and to plan response efforts. Twitter as successful real time micro-blogging social media, is increasingly used to improve resilience during extreme weather events/emergency management situations, including earthquake. It being used during crises by communicating potential risks and their impacts by informing agencies and officials. The geographical location information of such events are vital to rescue people in danger, or need assistance. However, only few messages contains there native geographical coordinates (GPS). So identifying location is a real challenge with Twitter data during critical situations. Identification of Tweets and their precise location are still inaccurate. In this work, we propose to use semi-supervised technique to utilize unlabeled data, which is often abundant at the onset of a crisis event, along with fewer labeled data. Specifically, we adopt an iterative Random Forest fitting-prediction framework to learn the semi-supervised model.**

***Index Terms*—Random forest, location prediction, location based-sensing system, Social networks, Twitter**

With the widespread use of mobile internet, hundreds of millions of people are spending countess hours on social medial to share, communicate, interact, and comments about an event that they have witnessed or heard about. The citizen participation in disseminating information during last years demonstrates the growing power of citizen influence on real life events [1]. This ”new media” is becoming one of the most significant channel for information contribution, dissemination and consumption which defines a new *citizen journalism concept* [2]. In a sense, this makes virtually every citizen a potential creator and user of information which can be used to evaluate the extent of a specific event. During an extreme event, individuals use social media to communicate, self-organize, manage, and mitigate risks (crisis-related communications) but also to make-sense of the event (commentary-related communications) [3]. The use of social media in emergencies has become a very big research field, sometimes summarized under the term of *crisis informatics* [4].

With approximately 336 millions worldwide active users per month posting (4th quarter 2018) a combined 500 million messages per day [5], Twitter is a successful micro-blogging platform where users write and share about topics which are catching their interest on the moment. These exchanges support people in times of crisis, and improve situation awareness of specific events, particularly in mass emergencies [6], such as weather events [7]–[9] and earthquakes [10]–[12]. Olteanu et al. investigated several crises in a systematic manner (information types, sources and their temporal distribution) and measured the prevalence of different types of Twitter messages under different types of crisis situations [13]. They showed that the average prevalence of sources as follows: **42% traditional or internet media**, **38% outsiders** (information originating from individuals that are not personally involved/affected by the event), **9% eyewitness accounts** (information originating from eyewitnesses of the event or of response/recovery operations, or from their family, friends, neighbors, etc), **5% government**, **4% NGOs** (Non-Governmental Organization) and **2% businesses**.

Two main features have been fundamental in Twitter success: the shortness of Tweets and the velocity of information transmission and of flows. Since 2017, Twitter increased the Tweet character number from 140 to 280-characters limit Tweets [14], [15]. A Tweet (and reTweet) is more than a short message, it comes bundled with a relatively rich set of metadata.

Twitter messages provide timely and fine-grained information about any kind of event. While these applications have been proven beneficial, the original location data recovery or the ability to effectively estimate the Tweets location has even more immense value. However, very few percentage of Tweets are geo-tagged in some way; for instance according to Cheng et al. only 0.42% Tweets have a native location coordinates:

This work received financial support from the MAIF fondation.

location data in the form of latitude and longitude [16], only 0.85 % are found in the research study conducted in [17] and about 3.17% Tweets are geo-tagged according to [18]'s study. This information reveals that Twitter has limited applicability as a location based-sensing system. In this context, accurately identifying from where a message originated from remains a challenge.

Location information on Twitter is available from two different sources:

- *Geotagging information*: users can optionally choose to provide location information for the Tweets they publish. The geographical latitude and longitude of the Tweet.
- *Metadata in the user's profile*: user location can be extracted from the location field in the user's profile. The information in the location field itself can be extracted using the APIs.

When the users switch on their geo-tagging, the information about the Tweet localization can be highly accurate, especially in the case when the Tweet is published using a smartphone with GPS capabilities.

For event analysis based on social network, one can consider Twitter as social sensor. Then we can implement spatial analysis of events by analyzing the social sensors. In this paper, we investigate locating a specified *earthquake* event crawled from Twitter. The problem of Tweets location inference can be generally formalized as a prediction problem. For each Tweet, we aim to predict where this Tweets comes from. In this work, we do not predict location at the country or city level; however, our approach concerns higher granularities: GPS points.

I. The Barcelonette earthquake

The earthquake of April 7, 2014 occurred at 21:27 local time (19:27 GMT) in the French region of Alpes-de-Haute-provence not far from the city of Barcelonnette. With a magnitude of 4.8, this moderate earthquake fortunately caused only small damages in a mountainous and relatively sparsely populated epicentral area. However, this earthquake is the largest earthquake in France since the appearance of Twitter, and its ground motions have been widely felt throughout southeast of France.

Today estimated at nearly 10.3 million, the number of French Twitter users was evaluated in 2014 at around 4.5 million, with a daily use still relatively undeveloped at the time of the earthquake. The occurrence of the earthquake, however, was manifested by a sharp peak of activity of Tweets mentioning keywords from the French lexical field related to earthquakes (see Chapter II-C for details on the collection of data).

A quick look at Figure 1 makes easy to highlight a significant increase in the number of these Tweets as of April 7, 2014 at 21:27 - the exact time of occurrence of the earthquake - with a Peak of 424 Tweets per minute reached at 21:29. This peak corresponds to a multiplication of a factor of more than 1400 compared to the average observed before the earthquake. In the first two hours after the earthquake, 8996 Tweets were collected, including 471 with GPS geolocation, corresponding to a particularly high ratio of 5.2% geolocated Tweets (but still not enough for fine cartography purposes).

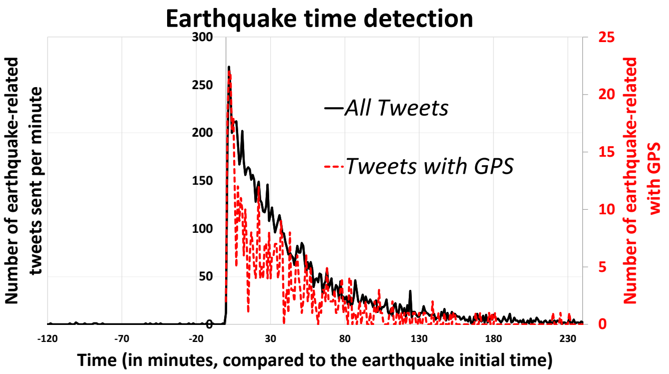

Fig. 1. Evolution of the number of Tweets collected every minute before and after the occurrence of the earthquake: all Tweets in black and dashed red line corresponds to those containing GPS coordinates.

II. Methodology

Figure 2 outlines the design and methodological scheme of the proposed method. This methodology draws on three main components: data pre-processing and labeling **(A)**, Spatial data filtering **(B)**, location geo-inference using random forest fitting-prediction and spatial interpolation using kernel probability density **(C)**.

The data pre-processing and labeling phase (box **(A)**) deals with the data collection, sampling and cleaning process. Proper data pre-processing is needed in order to use these Tweets for data labeling process. In the first step, non-French Tweets are filtered out from the raw database. A number of steps were used to clean the Tweets for this study. Since we were focusing on the textual content of the Tweets, the internet links were ignored. Details on labeling process is presented in II-C.

Following the pre-processing and labeling data box, the related event sample Tweets go towards the location feature extraction (box **(B)**). which will be explained in the section II-B. The data understanding phase needs any specific location elements that can have value for the analysis, in order to align each Tweet with the finest granular location. We obtained a bounding box in terms of latitude and longitude for all cities and states location names using **GeoNames**[1] API services.

The core function of the proposed method is the location inference using Random Forest [19], [20] successive fitting-prediction. The last step consists of an interpolation technique using a weighted scheme by Gaussian kernel of spatial probability density (box **(C)**). The steps of the box (**(C)**) is detailed in the next section.

A. The use of Twitter

Twitter has evolved as a popular micro-blogging website and consequently it is considered as very important source of

[1] http://www.geonames.org/

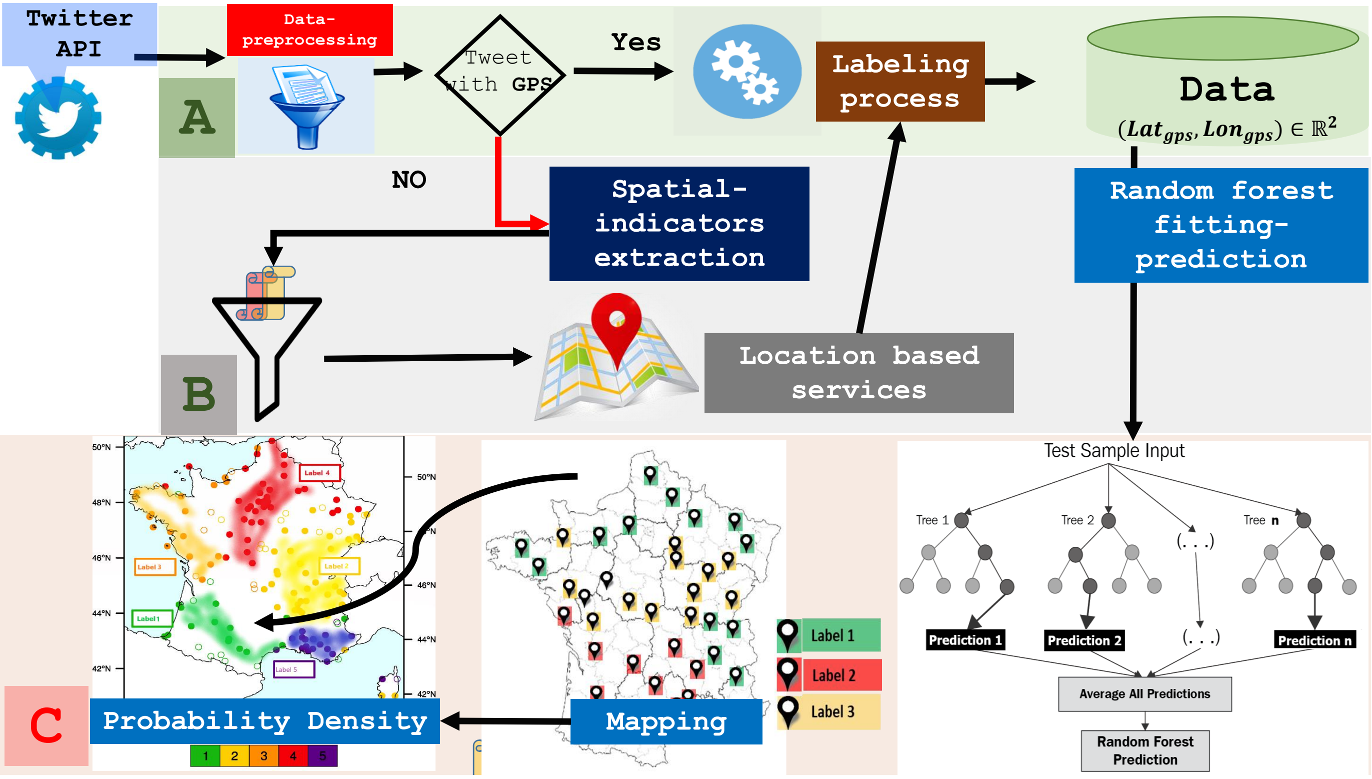

Fig. 2. Overview of the methodology used in the study

information. The rise of mobile internet users has significantly increased the number of Twitter users and provides an efficient medium for instant dissemination and consumption of information. The power of Twitter lies in its interactivity and its ability to amplify the reach of content.

To make the platform more flexible, Twitter has adopted topic suggestions, and user's mentions and provides different ways for users to interact by referencing each other in posted messages. Topics are grouped in Twitter using Hashtags, which is any keyword preceded by a hash sign '#' (eg.**#tremblement_de_terre**). To create a mention or a reply link to the referenced user's account, one can use a handle or place the "@" sign before a user name. Users can forward or re-Tweet someone else's Tweet to their followers, by using the RT prefix before the user name that originated the message.

Depending on the level of used permission authentication, the Streaming API allows the collection of some of published Tweets. Subsets of public status descriptions can be retrieved based on user-defined criteria in JavaScript Object Notation (JSON) formatted data.

B. Geographic location elements in Twitter data

Twitter API returns a JSON object for each Tweet, this is a common data exchange format consisting of a collection of key-value pairs. The JSON object contains Tweet content and various meta-data which may contain location references, there are location-specific elements that can have values of different types [15]. Some of these location elements in Tweets meta-data are used in this study:

- **`"Tweet⟶coordinates"`:**corresponds the exact Geographical coordinates provided in [LONG, LAT] order. The **"geo"** element provides the same information which has the reverse [LAT, LONG] order [15];
- **`"Tweet⟶place"`:**indicates that the Tweet is associated with a place, but not necessarily from this place. The user could attach a city name of the neighborhood of their choice to a Tweet. When present, Tweets bound with place are likely to be from within or around the place. These include entries such as the country and city associated with the place, as well as geographic coordinates.
- **`"Tweet⟶user ⟶geo-enabled"`:**indicates whether a user has ever chosen to share any location information. This field is boolean (TRUE or FALSE) and shows the case when users have agreed to turn on the location services at least once.
- **`"Tweet⟶user ⟶location"`:**defines the location for **`user`**'s account profile. This field might be filled with unexpected entries, not necessarily a compatible place with gazetteer location names database, i.e the users may lie or provide nonsensical locations. If the field is filled correctly , the locations are mostly static, corresponding to the user's primary location rather than the location at the time of the message posting, which may be different if the user is traveling.

All these fields do not necessarily contain a value, enabling the users to maintain some level of privacy and anonymity. After combining spatial indicators, we associate each recognized place toponym with a list of geographic interpretations via name lookup into a gazetteer (a geographical index). As *gazetteer*, we used **GeoNames** which is a database of geographic locations and associated meta-data that contains more than 10 million entries about spatial entities in different languages. This includes countries, cities as well as building and street names.

C. Data collection, pre-processing and labeling

In order to train and validate our model, sufficient Tweets related to an event are needed. In this study, we focus on Barcelonnette earthquake occurred on April 7th 2014. Because the earthquake event occurred before the beginning of our project, we used services of SIFTER (https://sifter.texifter.com/) which allowed until September 2018 the purchase of data from Twitter for past periods: we thus obtained a dataset of 29k tweets (of which 687 with available GPS geolocation) on the day of April 7, 2014, corresponding to a keywords based research using a corpus exploiting the lexical field of earthquakes (in French: ”*séisme*”, ”*tremblement de terre*”, ”*magnitude*”, ”*Barcelonnette*”, etc.). The exact query used is ”*seisme* OR *seismes* OR *séisme* OR *séismes* OR *tremblement de terre* OR *tremblements de terre* OR *magnitude* OR *terre tremble*”. Most of the collected Tweets were in French language. The Tweet text is labeled on three labels:

- *Témoin* (English↪ ***witness***): when the text talks about the event and its author claim to have seen or felt the consequences of the event.
- *Informatif* (English↪ ***informative***): when the Tweet informs us about the event, describes it or refers the consequences of the event, but it is difficult to say whether its author is a witness or not.
- *Hors-sujet* (English ↪ ***off-topic***): all other captured Tweets

III. Location estimation using Random Forest predictions

A. Random Forest estimation and prediction

Random Forest (RF) is an ensemble method in machine learning which involves construction (growing) of multiple Decision Trees (DTs) via bootstrap aggregation [19]–[22]. This is accomplished through the use of bagging and a classification and regression tree (CART) learning algorithm in order to build a large collection of “de-correlated” decision trees. The general framework in RF learning is non-parametric regression estimation. We assume we are given a training sample $\mathcal{D}_n = (\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)$ in which an input random vector is nonparametric regression estimation, in which an input random vector $\boldsymbol{X} \in [0,1]^p$ is observed. The goal is to predict the square integrable random response $Y \in \mathbb{R}$ by estimating the regression function $h(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$.

The goal is to use the data set $\mathcal{D}_n$ to construct an estimate $h_n : [0,1]^p \to \mathbb{R}$ of the function h. In this respect, a random forest is a predictor consisting of a collection of B randomized regression trees. For the j-th tree in the family, the predicted value at the query point $\boldsymbol{x}$ is denoted by $h_n(\boldsymbol{x}; \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \Theta_2, \ldots, \Theta_B$ are independent random variables, distributed the same as a generic random variable Θ and independent of $\mathcal{D}_n$. Note that the trees are combined to form the forest estimate:

$$h_{B,n}(\boldsymbol{x}; \Theta_1, \Theta_2, \ldots, \Theta_B, \mathcal{D}_n) = \frac{1}{B} \sum_{j=1}^{B} h_n(\boldsymbol{x}; \Theta_j, \mathcal{D}_n). \tag{1}$$

We now provide in the algorithm 1 the basic framework Random Forest based on individual trees.

```
Algorithm 1: Random forests basic framework
  Inputs : $\boldsymbol{X}$, $\mathcal{D}_n$and the number of tree $B$ s
  For $k = 1, 2, \ldots, B$ do
     • Draw a bootstrap sample in $\mathcal{D}_n$
     • Construct a CART tree on this bootstrap sample, each
       cutoff is selected by minimizing the cost function of
       CART over a set of $m$ variables randomly selected from
       the $p$. We note $h_k(\boldsymbol{x}; \Theta_j, \mathcal{D}_k)$ constructed Tree.
  End For
  Output : The estimation
  $h(\boldsymbol{x}) = \frac{1}{B}\sum_{k=1}^{B} h_k(\boldsymbol{x}; \Theta_k, \mathcal{D}_k)$
```

B. Random Forest for Tweets location

1) Basic considerations: The main idea of this study is to consider the geo-location inference as the analysis of incomplete multivariate data prediction problem for missing data. This study uses labels and coordinates as data input in the model. We address the geo-location inference problem using an iterative fitting-prediction scheme by training a machine learning model on observed values in a first step, followed by predicting the missing values (latitude and longitude) and then proceeding interactively.

We assume $\mathcal{D} = (\boldsymbol{Z}_i, X_{lab}, X_1, X_2, \ldots, X_p)$ to be a $n \times (p+3)$-dimensional data matrix with some missing geographical data $\boldsymbol{Z}_i$. Spatial locations are represented by latitude and longitude: $\boldsymbol{Z}_i = (X_i, Y_i)$ and Tweets labels are noted by X_{lab}. Other attributes of Tweets are represented in $(X_1, X_2, \ldots, X_p)$. By using RF, the Tweet label as response variable for training the forest is required, for this we directly predict the geographical missing values ($\boldsymbol{Z}_i$) using an RF trained on the observed parts of the dataset. As developed in [23], location inference is treated as a non-parametric missing value imputation. We recommend Stekhoven and Bühlmann’s work [23] for more details about computational efficiency.

For vector $\boldsymbol{Z}_s$ including missing values at entries $\boldsymbol{i}_{\text{NA}}^{(s)} \subseteq \{1, \ldots, n\}$, $\mathcal{D}$ can therefore be separated into four sub groups: *i)* Tweets with native GPS of $\boldsymbol{Z}_s$ (denoted $\boldsymbol{y}_{\text{obs}}^{(s)}$); *(ii)* new

Tweets without native GPS, considered as missing data (noted $\boldsymbol{y}_{\text{NA}}^{(s)}$), *(iii)* variables at entries $\{1,\ldots,n\} \setminus \boldsymbol{i}_{\text{NA}}^{(s)}$ other than $\boldsymbol{Z}_s$ (denoted $\boldsymbol{x}_{\text{obs}}^{(s)}$) and *(iv)* variables at entries $\boldsymbol{i}_{\text{NA}}^{(s)}$ other than $\boldsymbol{Z}_s$(denoted $\boldsymbol{x}_{\text{mis}}^{(s)}$).

2) The main procedure : The procedure starts initialization by using any method to full the data, for instance it might be possible to use mean imputation or another imputation method. For each variable (latitude and longitude) of $\boldsymbol{Z}_s$, the location Tweets without native GPS are predicted by iterative fitting-predicting an RF using the four part of $\mathcal{D}$ ($\boldsymbol{y}_{\text{obs}}^{(s)}$, $\boldsymbol{x}_{\text{obs}}^{(s)}$, $\boldsymbol{y}_{\text{NA}}^{(s)}$, and $\boldsymbol{x}_{\text{obs}}^{(s)}$). The procedure is repeated until a stopping criterion ν is met: when the difference between the newly predicted and the previous one increases for the first time. The computed quantity is: $\frac{\sum_{j\in\boldsymbol{k}}\left(\mathcal{D}_{\text{new}}^{\text{imp}}-\mathcal{D}_{\text{old}}^{\text{imp}}\right)^2}{\sum_{j\in\boldsymbol{k}}\left(\mathcal{D}_{\text{new}}^{\text{imp}}\right)^2}$, $\boldsymbol{k}=\{1,2\}$.

The pseudo Algorithm 2 gives a representation of the procedure.

Algorithm 2: Geo-inference by iterative random forest fitting-prediction procedure

```
Inputs : $\mathcal{D}$ original data and stopping criterion $\nu$
Initialization of complete $\mathcal{D}$: compute mean column if missing;
$\boldsymbol{k}=\{1,2\}$ indices of latitude and longitude of $\boldsymbol{Z}$.
While not $\nu$ do
    $\mathcal{D}_{\text{old}}^{\text{imp}}$ ←store previously imputed matrix ;
    For $s$ in $\boldsymbol{k}$
        Fit a random forest model : $\boldsymbol{y}_{\text{obs}}^{(s)} \sim f\left(\boldsymbol{x}_{\text{obs}}^{(s)}\right)$;
        Predict $\boldsymbol{y}_{\text{NA}}^{(s)}$ using $\boldsymbol{x}_{\text{NA}}^{(s)}$ : $\boldsymbol{y}_{\text{NA}}^{(s)} = \mathbb{E}\left(\boldsymbol{x}_{\text{NA}}^{(s)} \mid \boldsymbol{x}_{\text{obs}}^{(s)}\right)$;
        $\mathcal{D}_{\text{new}}^{\text{imp}}$ ← update matrix, using predicted $\boldsymbol{y}_{\text{NA}}^{(s)}$;
    End For
    update $\nu$;
End While
Output : The predicted values of $\boldsymbol{Z}_i = (X_i, Y_i)$ ;
```

C. Kernel density on estimated locations

Visualizing the density is useful for reporting results in a simple and understandable way. Kernel Density Estimation (KDE) is one of the most commonly used techniques for visualizing geographical data of a spatial process. KDE is a standard statistical technique to estimate a smooth probability density function. It has been extended from univariate distributions (on the real line) to multivariate distributions, including spatial and spatio-temporal models. In such estimation, we count the number of observations in the neighborhood of a given location: the closer the observation is, the greater its weight.

Spatial observations are based on spatial locations which are materialized by latitude and a longitude: $\boldsymbol{Z}_i = (X_i, Y_i)$. Based on a sample $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n\}$ the estimation of the density at unknown point $z = (x, y)$ is

$$\hat{f}_{\boldsymbol{H}}(z) = \frac{1}{n}\det(\boldsymbol{H})^{-1}\sum_{i=1}^{n}\mathcal{K}\left(\boldsymbol{H}^{-1}(z-\boldsymbol{Z}_i)\right), \qquad (2)$$

where $\mathcal{K}$ is some symmetric (centered) kernel function, and $\boldsymbol{H}$ a bandwidth parameter. In this work, we used Gaussian kernel.

IV. Results and discussion

Figure 3 presents the location estimation of an earthquake occurred in Barcelonnette. presents the predictions of latitude and longitude prediction according to the type of the Tweet: witness, informative and off topic. In order to simplify the map interpretation the results can be summarized as follow:

- The gradient color is used to estimate the probability density of observation of points over a geographical area. The inferred area with a higher probability is superimposed on the area of the event.
- the gradient color is a visual support of the contours in order to estimate the density of probability of observation of points over a geographical area.

The prediction model performs well in the witness region. One note that it becomes more difficult to make a good estimation in less Tweeted areas about the event, which correspond to less-populated areas. This could be attributed to the fact that the greater the number of witness "*sensors*", the more precise the estimation will be. Since the predictions of latitude and longitude are performed simultaneously, it could be that the model prediction may performs well in one variable compared to another.

Locating Tweets on a map - as well as a close reading of the messages - highlights a particularly marked cluster of activity along the French Riviera, particularly in Nice, in a densely populated area where the earthquake was widely felt by the population. In the North, the earthquake also gave rise to many Tweets from the city of Grenoble for similar reasons. More surprising, the activity peaks observed nearby cities of Lyon (located about 200 km north-west of the epicenter) and Marseilles (located about 170 km southwest of the epicenter): although the earthquake was weakly felt there, these peaks of activities can be explained not only because of the messages from witnesses, but also because of discussions about the earthquake risen after its announcement through Twitter user's timelines and continuous TV news channels. These comment messages also explain the well-marked peak activity observed in the Paris region.

The Figure 4 allows a comparison to be made between test Tweets (with a GPS location) and those inferred by the model.

V. Conclusion

The current research attempts to effectively use social media for location inference of Tweets. The very low volume of geotagged Tweets makes location inference a necessity and the accuracy of the location inference is crucial for accuracy of the earthquake location. We have outlined one location inference method here, which flows the data imputation principles. The main contribution of this study is to tackle the problem as a statistical learning approach for missing data.

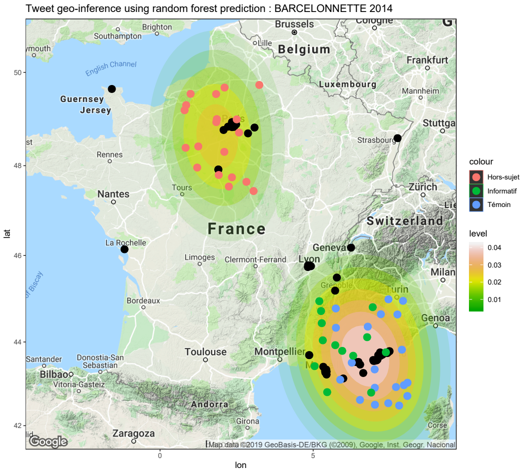

Fig. 3. Earthquake location estimation based on Tweets using latitude-longitude prediction. The black dots correspond to the actual location of the Tweets we are trying to find (test set). All points in color correspond to localizations inferred by the method and each color informs the Tweet label (Témoins ↪ witness; informatif ↪ informative; hors-sujet ↪ off-topic). The prediction are given according to the labels for 5 % of test data

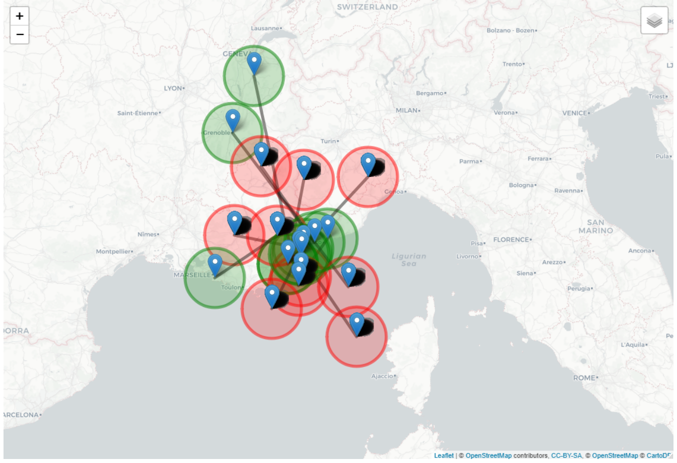

Fig. 4. Earthquake location estimation based on Tweets using latitude-longitude prediction. A snapshot of an interactive chart: the green circles represent the real location and the red ones correspond to the prediction locations.

The approach have bias, since the twitter user-base is not representative of the population in general.

For the future work, we will apply this result to the existing event detection systems to confirm that our result will improve their performance in terms of the event location estimation. There are at least one open issue needing further studying: both Random Forest and kernel density methodologies suffer a so-called *"edge effect"* or *"border bias"*. So how to improve the accuracy of our estimation at is worth studying in the future.

Acknowledgment

This work received financial support from the MAIF fondation.

References

[1] D. Villatoro and J. Nin, "Citizens sensor networks," in *International Workshop on Citizen in Sensor Networks*. Springer, 2012, pp. 1–5.

[2] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[3] S. Stieglitz, D. Bunker, M. Mirbabaie, and C. Ehnis, "Sense-making in social media during extreme events," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 4–15, 2018.

[4] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," *International journal of emergency management*, vol. 6, no. 3-4, pp. 248–260, 2009.

[5] Statista, "Number of monthly active twitter users worldwide from 1st quarter 2010 to 4th quarter 2017 (in millions)," Jan. 2018. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[6] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[7] K. Tamura and T. Ichimura, "Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2079–2084.

[8] R. Arthur, C. A. Boulton, H. Shotton, and H. T. Williams, "Social sensing of floods in the uk," *PloS one*, vol. 13, no. 1, p. e0189327, 2018.

[9] F. Laylavi, A. Rajabifard, and M. Kalantari, "A multi-element approach to location inference of twitter: A case for emergency response," *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, p. 56, 2016.

[10] O. Ozdikis, H. Oguztuzun, and P. Karagoz, "Evidential location estimation for events detected in twitter," in *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM, 2013, pp. 9–16.

[11] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "# earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, 2013.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[13] A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 994–1009.

[14] A. Rosen and I. Ihara, "Giving you more characters to express yourself," *Twitter Blog*, vol. 26, 2017.

[15] Twitter. (2018) Tutorials filtering tweets by location. https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.html. Accessed: 2018-10-10.

[16] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[17] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana, "Knowing the tweeters: Deriving sociologically relevant demographics from twitter," *Sociological research online*, vol. 18, no. 3, pp. 1–11, 2013.

[18] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." in *ICWSM*, 2013.

[19] L. Breiman *et al.*, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] ——, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[22] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[23] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011.