



Spatial patterns of the French rail strikes from social networks using weighted k-nearest neighbour

Rachid Ouaret, B. Birregah, Omar Jaafor

► To cite this version:

Rachid Ouaret, B. Birregah, Omar Jaafor. Spatial patterns of the French rail strikes from social networks using weighted k-nearest neighbour. International Journal of Social Network Mining, 2020, 3 (1), pp.52. 10.1504/IJSNM.2020.105745 . hal-03296912

HAL Id: hal-03296912

<https://hal.science/hal-03296912>


Submitted on 22 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spatial patterns of the French rail strikes from social networks using Weighted k -Nearest Neighbour


This paper was submitted to IJSNM. The published version can be found in IJSNM 2020 Vol.3 No.1, pp.52 - 76. Click on the link below:

DOI  10.1504/IJSNM.2020.105745

Rachid OUARET*

Department Operational Research and Applied Statistics (ROSAS),
University of Technology of Troyes, Troyes, France

E-mail: rachid.ouaret@utt.fr

Update contact details  : rachid.ouaret@toulouse-inp.fr

*Corresponding author

Babiga BIRREGAH

Department Operational Research and Applied Statistics (ROSAS),
University of Technology of Troyes, Troyes, France

E-mail: babiga.birregah@utt.fr

Omar JAAFOR

Department Operational Research and Applied Statistics (ROSAS),
University of Technology of Troyes, Troyes, France

E-mail: omar.jaafor@utt.fr

Abstract: The information analysis provided by millions of social network users is one of the most important sources of information yielding interesting insights of spatial patterns about Socio-Political events. During the recent French National Railway strikes (from April to June), Twitter was used as platform where people expressed their opinions, with millions of “SNCF” (French National Railway Company) tweets posted over the strike period. In this paper, we have discussed a methodology which allows the utilization and interpretation of Twitter data to determine spatial patterns over French territory. The identification of a geographic strike landscape is achieved through spatial interpolation using Weighted k -Nearest Neighbour. This study shows the benefits of geo-statistical learning for extracting sentiment polarities of social events across France.

Keywords: Weighted k -Nearest-Neighbor; spatial interpolation; Twitter; social networks; railway network; social polarities; sentiments analysis

Biographical notes: Dr. OUARET Rachid. received the M.S. degree in Information processing and data analysis from the Paris Sud University and

the Ph.D. degree from the University Paris-Est, France, in 2016. He worked as teaching and research assistant at the University of Eastern Paris (UPEC) for Two years. Since November 2017, he is working as post-doctoral researcher on social network data analysis and mining at University of Technology of Troyes (Department of Operational Research and Applied Statistics). His research interests lie on the time series analysis and forecasting, dynamical systems and spatial statistics on social networks. He has been involved in several research projects on data analysis and served as a Referee for many journals and conferences.

Dr. BIRREGAH obtained a PhD in Applied Mathematics in 2007. He then joined the University of Technology of Troyes as a post-doctoral researcher and researched on the Surveillance, Safety and Security Group (3SGS). In September 2009, he joined the CNRS Joint Unit Charles Delaunay Institute (ICD) as an assistant professor. Babiga's research interests are on Data analysis for System security and resiliency with application in Cybersecurity and Crisis management. As member of ICD he has developed research activities in Resilience in the Interdisciplinary Group devoted to risk assessment and management technologies. He has participated in several projects, funded by regional and national agencies in Big Data. Also, he has worked with several companies in Railway and Security. From 2013 to 2016, Babiga was in charge of the DU "Criminal Operational Analysis" (co-labeled UTT and National Gendarmerie). Since 2016 he has been the management lead of the Specialized Master® Expert Big Analytics and Metrics of UTT. He is the author and contributor of several software systems dedicated to large scale graphs analysis and mass railway transportation systems.

Dr. JAAFOR Omar Received a Master degree in the security of information systems from University of Technology of Troyes, France in 2013 and a Master degree in computer science from the University of Jean Monnet in Saint Etienne, France in 2014. He receive a PhD in anomaly detection in social networks at University of Technology of Troyes. Since the start of his PhD in 2014, he has been performing research in anomaly detection, collective classification, sentiment analysis and unsupervised classification. He is currently a data science and machine learning consultant at Novagen Conseille, Paris.

1 Introduction

Since the rise of Social Network Services (SNS) in the early 2000s, and particularly Twitter towards the end of the decade, users of social network services have grown from the hundreds of millions to over one billion. Over the same period, social network data has become one of the most effective and accurate indicators of public opinion (Hridoy et al. 2015, Murphy et al. 2014). This is because several million people are connected to social network platforms at any time to share, communicate, interact, and comment about events that they have witnessed, interested in, or have heard about. Contrary to traditional surveys, social media data offers several advantages, which include the cost efficiency at which the data can be collected, analysed, and disseminated.

With approximately 335 million active users worldwide per month (2nd quarter 2018), posting a combined 500 million messages per day (Statista 2018), Twitter is a successful micro-blogging platform where users discuss topics that catch their interest at the moment. Using Murphy's terms: Twitter is becoming the microphone of the masses (Murphy et al. 2014), which is altering news production and consumption. Political opinions about special

87 events on Twitter partly express political polarization, indicating that the content of Twitter
88 messages plausibly reflects the off-line political landscape (Tumasjan et al. 2010).

89 The public opinion expressed in Twitter is used as an indicator of economic power.
90 In an early work trying to predict stock market indicators by analysing Twitter posts, the
91 rate of emotion on Twitter (hope, fear, joy) was proportional to the evolution of stock
92 market indices (Zhang et al. 2011). Several case studies seem to show a strong correlation
93 between the analysis of tweets and the evolution of economic indices (Bollen et al. 2011,
94 Ranco et al. 2015). Public opinion has always been an important source of information for
95 most politicians during the decision-making process (Gaber 2017, Tumasjan et al. 2011).
96 Understanding the political representativeness of Twitter users is feasible through analysing
97 social events/movements (Mercea & Yilmaz 2018). From this perspective, the fundamental
98 battle in social debates is the battle over opinions and feelings of the people. The way people
99 think and feel determines the norms and values on which societies are constructed (Castells
100 2015).

101 This article attempts to deepen the study of communication in large social movements
102 through an examination of Twitter usage, particularly strike related messages. We will focus
103 on the latest French railway strikes.

104 On March 14th 2018, the French Minister of Transport presented the draft law on the
105 railway reform to the Council of Ministers. In response to this proposed legislation, a strike
106 was announced by the railway worker's unions of the French National Railway Company,
107 "Société Nationale des Chemins de fer Français (SNCF)". In this socio-political situation,
108 Twitter offered a broad platform for people to express their opinions and feelings with
109 millions of "SNCF" tweets posted during and after the strike period. Recourse to the public
110 sphere is a dominant feature of social protests because the success of a strike depends on
111 the growing weight of the media and public opinion.

112 In many ways, the French rail system can be considered as one of most successful in
113 the world, with a vast well-developed network, a dynamic regional transport, as well as a
114 good, coverage of the territory.

115 The rail system plays a major role in the mobility of the French people, meeting the
116 needs of travellers both for mass departures on vacation and for their daily or professional
117 trips. This rail system is concentrated in Ile-de-France (Parisian Region), and spread over
118 the rest of the country. It plays a major role in the mobility of goods, and many activities
119 (including agriculture, steel and automotive industries, ...) are dependent for the delivery
120 of their production. However, the French rail system is generating growing dissatisfaction,
121 from users, travellers, shippers, railway companies and transport authorities. The poor state
122 of the network, with its failures and the many works, does not explain everything. To those
123 who long for SNCF reform : the regularity is insufficient, security seems to be compromised,
124 supply is declining, even though public funds devoted to rail have never been so high.

125 In the unions, the worker's point view can be summarized in four points:

126 (i) Reform of their status: new employees will not benefit from status which grants many
127 economic advantages and which symbolizes the historic social achievements of the
128 railwaymen.

129 (ii) The transformation of the SNCF into a public limited company provided for in the bill
130 would, according to the unions, lead to future privatization and is also at the centre of
131 their concerns.

(iii) Railway workers are worried about possible closures of small regional lines. The profitability of 9,000 kilometres of these lines was called into question by the Spinetta report, submitted to the government in mid-February-2018.

(iv) Trade unionists believe that the debt of the SNCF, which represents 45 billion euro, weighs too much on the company, while it is the state that should be responsible.

In France, where the state is an important actor in many industries, strikes are frequently used as a form of political action to exert pressure on the government. These strikes reveal the enormous strength of the "class cleavage" in France (Hanspeter et al. 2015).

To better meet this new reality, on March 15, 2018, the French railway workers' unions opted for a strong social movement, a strike at a rate of "two days out of five" from April to June against the reform of the "*Société Nationale des Chemins de fer Français (SNCF)*" that the governments intends to conduct quickly by accelerated legislation. The reform advocates bringing the SNCF company up to European standards, ending its social model and closing unprofitable lines. Reforming the SNCF is a highly controversial and sensitive subject as the national society is closely identified with the history of modern France.

It is in the interests of large public institutions to be able to retrieve all the information related to strikes and their development over time. Making use of this information flow, one can distinguish what is important in order to provide better understanding of strikes. This raises an obvious question for the French government: How does the ubiquity of social media affect public opinion on rail reform?

We can understand the pressure exerted on the government by strikes by analysing the spatial clusters of "support" and/or "against" the strikes over the entire French territory. A full picture of the pressure zones could yield interesting insights in anticipating the future strikes as well as predicting their political orientations. Supervised learning and sentiment analysis techniques are an effective means for discovering French public opinion on the SNCF strikes.

This article starts by describing the data about SNCF strikes on Twitter. It then moves to presenting a method that allows the characterization and spatial interpolation of messages published in an on-line social networks. This article concludes with a set of experiments detailing the spatial patterns and sentiments of the French population concerning the SNCF strikes.

2 Methodology

2.1 Trends of the SNCF related words over 5 years

The primary data source that we used for the identification of spatial patterns related to railway strikes was Twitter data. To understand the SNCF's social movements on the web better, we used Google Trends.

Google Trends Engine provides an index of the relative volume of search queries conducted through Google. Note that the Google trends adjust search data to make comparisons between terms easier. The trend does not indicate an absolute number of searches but a proportion between 0 and 100, where 100 represents the maximum amount of the term used in the defined period and location in order to compare relative popularity. This tool allows us to estimate the impact of social movements on internet user activity, on

the social networks as well as traditional media. Figure 1 shows the relative popularity in France of a search query: "grève SNCF" (SNCF strike), "Grève" (strike) and "SNCF".

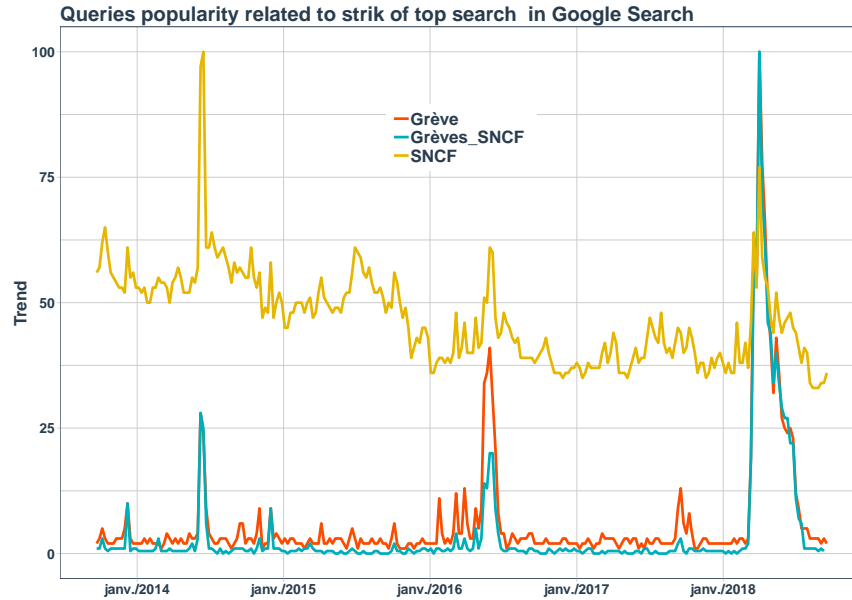


Figure 1: Strike related queries popularity top search in Google Search Engine during the last 5 years

The two time series associated with the key words "SNCF strike" and "strike" are highly correlated. Whereas there are very few peaks of the curve "strike" having a correspondence with the series "strike SNCF". This result shows that most of the strikes conducted by the various social movements are the result of strikes at the SNCF.

2.2 The use of Twitter

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

Twitter allows people to create profiles, communicate, and connect with other people on the service. A user can follow any other user without requiring approval or a reciprocal connection from the followed users. The follower network describes the "virtual" social relationship on Twitter and can be conceptualized as a directed social network (Brzozowski & Romero 2011).

Messages, called tweets can be posted on Twitter through various communication services, that allow portability, immediacy and ease of use. 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. **#JeSoutiensLesCheminots**). 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 retweet someone else's tweet to their followers, by using the RT prefix before the user name that originated the message.

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 (Rosen & Ihara 2017, Twitter 2018). A tweet (and retweet) is more than a short message, it comes bundled with a relatively rich set of metadata.

Depending on the level of used permission authentication, the Streaming API allows the collection of up to 1 % of all published tweets. Subsets of public status descriptions can be retrieved based on user-defined criteria in JavaScript Object Notation (JSON) formatted data which is a lightweight text-based data exchange format.

Using the Streaming API results to perform data analysis can raise issues concerning the validity of the data due to the quality of the sample and any eventual bias. A number of methods are proposed by different researchers to analyse Twitter data for varied purposes.

2.3 *Data collection and processing*

The dataset used in the scope of this paper was collected between April 4, 2018, and June 30, 2018, using the streaming API. Only the collection of tweets produced using the "SNCF" keyword and related strike movements of railway workers were retrieved. All data was collected according to Twitter's terms of service and privacy conditions. We filtered the stream of tweets keeping only those which were published in France (986K tweets). This data represents a good picture of the strike happening in France over this period. However, the pre-processing step obliges us to not include most of tweets, as the re-tweets are removed from the study.

2.3.1 *The labelling process*

In order to calibrate and validate our model, sufficient tweets related to an event are needed. No restrictions on location or user was implemented. However, query terms were limited to topics related to the SNCF strikes in France. Building models to classify data according to a predefined coding scheme is an essential task in digital social research, used for the purpose of understanding social interactions, beliefs and emotions. In this research, once Twitter data was collected, we first used Natural Language Processing and after that we built a supervised machine learning model to identify spatial patterns and to explore socio-geographical polarity about the the rail strikes sentiment across France.

To complete this subjective task using large-scale data analytic, which is essential for the volumes of data produced, we used machine classifiers to learn the features of tweets that are indicative of the class they belong to.

In the first step, we manually labelled 654 tweets on five labels *Support (Favour)*, *Oppose (Against)*, *informative*, *Neutral* and *Irrelevant*. We first used these labels to capture the finest spatial polarity and afterwards, reduced it to three main categories *Support (Favour)*, *Oppose (Against)* and *Irrelevant*. The Table 1 shows the manual labelling used in the dataset according to the opinion and sentiment conveyed by the text.

The next step involves building a sentiment analysis model to extrapolate the opinion of 4555 tweets: we used the text in the retrieved tweets to detect their opinion with regards

Table 1 Manual Labeling according to the sentiment / opinion conveyed by the Twitter-text .

Labeled tweets	Sentiment /opinion orientation	Examples: original French tweets	English version of the tweets
Favour ⇒ support the strike movement	the twitter-text is in favor strike and the social movement	<i>“#JeSoutiensLaGrèveDesCheminots Je suis heureux de voir qu’il y a encore des gens courageux pour lutter contre la politique injuste et dévastatrice de #Macron. Prenons tous exemple sur les #cheminots !”</i>	<i>“#ISupportTheRailwaymenStrike I am happy to see that there are still brave people fighting against #Macron’s unfair and devastating policy. Let’s take an example of the #railwaymen!”</i>
Against ⇒ oppose the “cheminots” movements	the twitter-text reveals an opinion against or openly opposing the strike	<i>“Enfin la fin des privilèges pour ces feignasses ! Le pire avec cette grève c’est qu’ils scient la branche sur laquelle ils sont ? #JeSoutiensLaGreveDesCheminots”</i>	<i>“Finally the end of the privileges for these lazy men! The worst thing about this strike is that they are scouting the branch they are on?#ISupportTheRailwaymenStrike”</i>
Neutral / Informative ⇒ related to strike context	the twitter-text gives either information on the strike or does not clearly position on the social movement	<i>“BREAKING NEWS : une ministre du gouvernement -monarque-sous- influencé - va s’exprimer sur la grève des cheminots : à vous madame la ministre.....heu ? Non ? Le trac, sans doute ????”https://t.co/AiE0iJB5pH</i>	<i>“BREAKING NEWS :a minister of the monarch-under-influenced government will speak about the railway workers’ strike: to you, minister ... uh? No ? The fright, probably ????”https://t.co/AiE0iJB5pH</i>
Irrelevant ⇒ they had not relationship to strike event	the twitter-text is irrelevant for the study	<i>“Merci de nous rappeler de cette forme de nostalgie positive qui nous raconte que leur story telling a bien fonctionné que de nos jours ne voulons plus être dupés...” https://t.co/LONI2nliA2</i>	<i>“Thank you for reminding us of this form of positive nostalgia that tells us that their story telling has worked well that nowadays do not want to be fooled anymore...” ...https://t.co/LONI2nliA2</i>

Table 2 Proportion of labels in manual and automatic cases

Proportions	Number of tweets	Favour /Support	Against / Oppose	Off Topic		
				Neutral	Informative	Irrelevant
Manual labeling	654	10.40%	17.1%	12%	14.20%	46.3%
Automatic labeling	4555	11.75%	19.67%		68.58%	

to the riots in addition to their sentiment. The tweets were modelled using a bag of words model. The dataset used to build the sentiment analysis model was retrieved from (Paroubek et al. 2018) who provide a corpus that contains 76K labeled tweets concerning the major SNCF strike in 2018.

The data sources come from **2018 DEFT text mining challenge** (*Analysis of tweets on transport on the Île-de-France*). The corpus consists of tweets in **French** that relate to transport in Île-de-France. It contains 76,732 tweets selected from 80,000 tweets annotated manually.

The model used is based on a simple logistic regression with extensive data preprocessing in order to detect symbols (... , !!! , etc.) that reflect the sentiment of a tweet. This simple model achieved a 90.72% accuracy using a 10 fold cross-validation on the trained data, which is high with respect to the difficulty of sentiment classification. The classification that was performed concerns the detection of the opinions of users with regards to the strike. We classified every tweet into the three following categories 1) supports the strike 2) opposes the strike and 3) neutral or unknown. We then used a multinomial Bayes algorithm to generate the model. The model obtained an average accuracy of 82% using a 10 fold validation on the training data. This is a satisfying result given the small number of tweets used in training.

Table 2 describes the data used in the analysis according to the labelling : manual and automatic scheme.

2.3.2 Geographic location elements in Twitter data

Twitter 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 (Twitter 2018). Some of these location elements in tweets meta-data are used in this study:

(i) **"Tweet→coordinates"**: corresponds to the geotagging which contains the exact Geographical coordinates provided in [LONG, LAT] order. The **"geo"** element provides the same information which has the reverse [LAT, LONG] order (Twitter 2018);

(ii) **"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 neighbourhood 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.

(iii) "**Tweet** \rightarrow **user** \rightarrow **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.

(iv) "**Tweet** \rightarrow **user** \rightarrow **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 travelling.

All these fields do not necessarily contain a value and can be left blank, 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**¹ 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. From 12k tweets in the pre-processed step, we obtained 4555 tweets associated to a place with longitude and latitude.

2.4 Method design and development

Figure 2 outlines the design and methodological scheme of the proposed method. This methodology draws on four main components: data pre-processing (A), Data preparation and understanding (B), Mapping and cell grid definition (C), Spatial interpolation (geostatistical approach) using weighted kernel k -NN (D).

The data pre-processing phase (box A) mainly deals with the data collection, sampling and cleaning process. Proper data pre-processing is needed in order to use these tweets for any meaningful purpose. In the very 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: the internet links (starting with <http://>) were deleted from the tweets. Removing internet links to photos, videos or news items in a tweet can result in loss of useful detailed information about the strike opinion polarity. However, since we were focusing on the textual content of the tweets, the internet links were ignored. In fact, we tokenize all tweets in the training dataset, which removes punctuation and other white- space. There are many accounts from automated tweets 'bots' and most of them are irrelevant and high likely to be designed for marketing purposes and have to be taken out of the analysis.

Following the pre-processing data component, the SNCF strike related events sample tweets go towards the location feature extraction and data labelling process (box B), which were explained in the previous section. 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 (**GN**) API services.

The next step consists of a territorial grid and assigning symbols and labels to the underlying feature geometry (box C). A higher number of grid cells (pixels) almost quadratically increases computing time. A lower number takes less time to compute, but yields a less continuous, more pixelated surface.

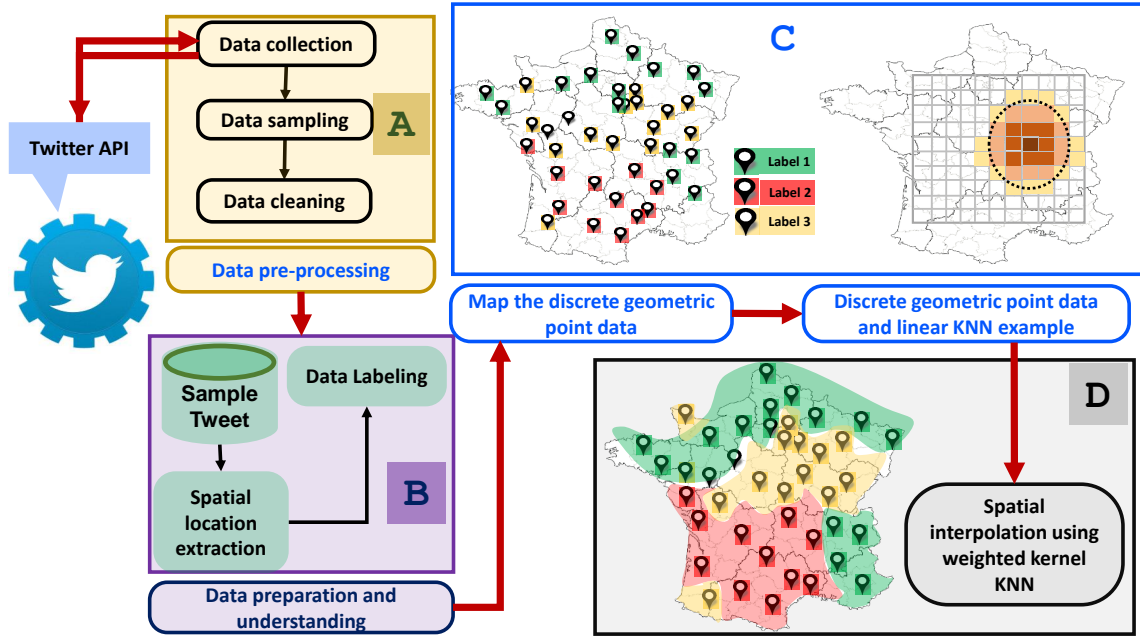


Figure 2: Overview of the methodology used in the study

318 The core function of the proposed method is the spatial interpolation technique using
 319 a weighted scheme by kernel functions of k -nearest neighbours for each cell grid (box D).
 320 The last step is detailed in the next section.

321 3 The main models

322 The originality of this research lies in the combination of a traditional kernel density estimate
 323 for k -nearest neighbours method and spatial features that describe the sentiment polarity
 324 about social movements and its trend using Twitter content. We defined spatial sectors
 325 (neighbourhoods) in metropolitan France by laying down evenly spaced cells 1000 meters
 326 on each side:

- 327 (i) Set how many grid cells (pixels) there should be the spatial neighbourhoods;
- 328 (ii) Create a regular grid with a certain size of cells (in meters) over the extent of the sample
 329 feature object. The simple feature object describes how objects in the real world can
 330 be represented by computer with an emphasis on the spatial geometry of the object.

331 Given the neighbourhood defined above, the problem becomes the estimation of the
 332 sentiment polarity values for each neighbourhood given the tweets posted by users. The
 333 goal of the combination method developed in this study is to interpolate a discrete geometric
 334 point data set to a continuous surface, represented by a regular grid (raster).

3.1 Outline of k -Nearest Neighbour (k -NN)

Since being introduced by (Fix & Hodges Jr 1951), the k -nearest neighbour algorithm (k -NN) has been one of the most well-known non-parametric supervised learning algorithms. The K -NN is a direct extension of the Nearest Neighbour rule (NN) (Cover & Hart 1967).

The principle of the k -NN (NN for $k = 1$) rule can be summarized as follows: let $T = \{(x_i, y_i)\}_{i=1}^N$ learning set of observed data provided with the distance d , where $x_i \in \mathbb{R}^m$ is training vector in the m -dimensional feature space, and $y_i \in \{1, \dots, c\}$ is the corresponding class label (a class membership). Given a query x^q , its unknown class y^q is assigned using an arranged set in an increasing order of the query T^q in terms of an arbitrary distance function $d(x^q, x_i^{NN})$ between x^q and x_i^{NN} , $i \in \{1, \dots, k\}$. That is x_1^{NN} is the nearest Neighbour and x_k^{NN} is the farthest of the k nearest neighbours.

Let us denote this sample set of the k similar labelled target neighbours for the query x^q as $T^q = \{(x_i^{NN}, y_i^{NN})\}_{i=1}^k$, the class label prediction of the query y^q , can be obtained by the majority voting of its nearest neighbours :

$$y^q = \arg \max_y \sum_{(x_i^{NN}, y_i^{NN}) \in T^q} \delta(y = y_i^{NN}). \quad (1)$$

where y_i^{NN} is the class label for the i -th nearest neighbour among its k neighbours and $\delta(\bullet)$ is the Dirac delta function: $\delta(y = y_i^{NN}) = 1$ if $y = y_i^{NN}$ and 0 otherwise.

3.2 Weighted k -Nearest Neighbour (wK -NN)

3.2.1 The main idea

The k -NN rule assumes that the neighbours' contributions are identical, regardless of the distance that separates the query from the neighbours. As an improvement to k -NN, Dudani introduced a distance-weighted k -NN rule (wK -NN) with the basic idea of weighting close neighbours more heavily, according to their distances to the query (Dudani 1976). More simply, the observations within the learning set, which are particularly close to the new observation, should get higher weights in the decision than those that are far away from the new observation. Figure 3 shows the prediction results using Weighted k -Nearest Neighbour: the sum of the weights of the neighbours from the blue circle class ($\sum_{i \in \bullet, i \leq k} w_i = 0.95$) is larger than that of the neighbours of the red square class ($\sum_{i \in \blacksquare, i \leq k} w_i = 0.45$).

With the Weighted k -Nearest Neighbour rule, each neighbour x_i^{NN} , $1 \leq i \leq k$, is equipped with a weight w_i , $1 \leq i \leq k$. The computation of weights can be defined by different methods. Dudani 1976 proposed Distance-weighted k -NN that assigns a weight w_i to the i th nearest neighbour as a function of $d(x^q, x_i^{NN})$. The weight for the i th nearest neighbour of the query x^q , as defined in (Dudani 1976), is as follows:

$$w_i = \begin{cases} 1 & , \text{if } d(x^q, x_k^{NN}) = d(x^q, x_1^{NN}) \\ \frac{d(x^q, x_k^{NN}) - d(x^q, x_i^{NN})}{d(x^q, x_k^{NN}) - d(x^q, x_1^{NN})} & , \text{otherwise} \end{cases} \quad (2)$$

The weights of neighbours with smaller distances are more heavier than ones with greater distances. The nearest neighbour gets a weight of 1 and the other neighbours' weights are

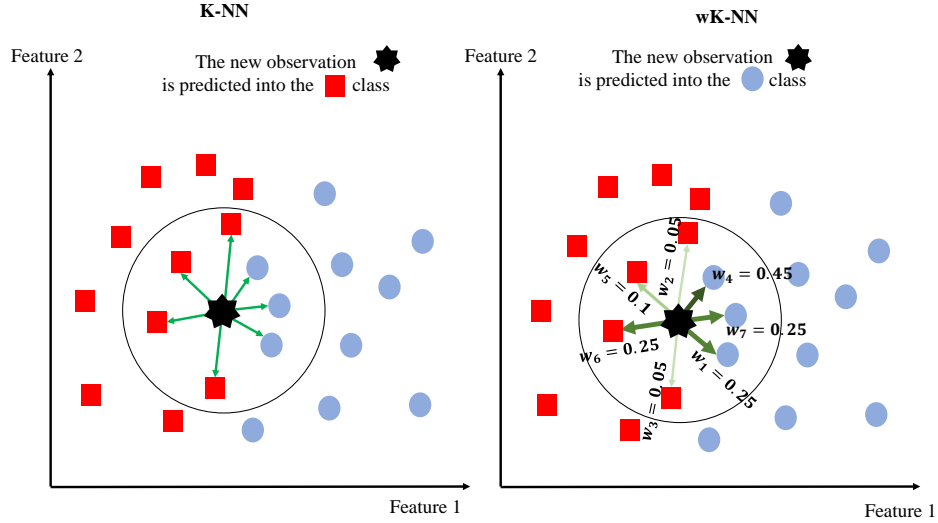


Figure 3: Example of k -Nearest Neighbour and Weighted k -Nearest Neighbour with $k = 7$. In the k -NN case, every neighbour counts in the same way for the final decision: all the arrows from the target to neighbours have the same thickness. Thus the prediction of the black star (new observation) is assigned to the red square class, which is the most frequent class in the neighbourhood. In the Weighted k -NN case, every neighbour has its own weight that influences the final decision (the the arrows have different thicknesses), the black star is then assigned to the blue circle class.

372 scaled linearly to the interval between 0 and 1 with weight of 0 for the furthest neighbour.
 373 The classification result of the query is made by the majority weighted voting:

$$374 \quad y^q = \arg \max_y \sum_{(x_i^{NN}, y_i^{NN}) \in T^q} w_i \times \delta(y = y_i^{NN}). \quad (3)$$

375 The theoretical framework of Dudani's proposition showed that, given an infinite set of
 376 training samples, the asymptotic error rate of unweighted k -NN is always lower than that of
 377 any weighted k -NN (Dudani 1976). This result was pioneered by Bailey and Jain (Bailey
 378 & Jain 1978). However, would not apply when the number of training samples is finite:
 379 when dealing with the realistic setting of finite samples, improvements using weighting are
 380 possible (MacLeod et al. 1987, Bicego & Loog 2016). weighting scheme in Equ. 2

381 3.2.2 Weighting scheme using Kernel functions

382 The weights can be estimated by using kernel functions $K(\bullet)$ of the distances d . According
 383 to the Hilbert-Schmidt theory, $K(d)$ can be an arbitrary symmetric function that satisfies
 384 the Mercer condition (Courant & Hilbert 1962). From empirical point of view, the choice
 385 of a special kernel is not crucial, apart from the rectangular kernel, that gives equal weights

to all neighbours (Hechenbichler & Schliep 2004). The kernel function is applied to the standardized distances obtained using the following formula:

$$D_i = D(x^q, x_i^{NN}) = \frac{d(x^q, x_i^{NN})}{d(x^q, x_{k+1}^{NN})} \text{ for } i = 1, \dots, k \quad (4)$$

In order to avoid weights of 0, a small positive constant ($\varepsilon > 0$) should be added to $d(x^q, x_{k+1}^{NN})$. By using kernel weighting, each new case is classified into the largest added weight as:

$$y^q = \arg \max_y \sum_{(x_i^{NN}, y_i^{NN}) \in T^q} K(D_i) \times \delta(y = y_i^{NN}). \quad (5)$$

The kernel weighting can be viewed as a generalization of k -NN and NN methods. That is, when using rectangular kernel, its result will be identical weights as k -NN. In the case when using $k = 1$, the results are similar as NN, independently of the kernel function. According to Hechenbichler and Schliep's remarks (Hechenbichler & Schliep 2004), the most important results of using kernel weighting is that the number of nearest neighbours, k is implicitly hidden in the weights.

Generally it can be said that the weighted scheme by kernel extends the basic k -NN method in two directions:

- the mode of the class probability distribution determines the voting of nearest neighbours
- it uses the median or the mean of that distribution, if the target variable shows an categorical/ordinal or even higher scale level.

4 Experiments and results

4.1 Choosing k and kernel parameters

Like most machine learning algorithms, the k in k -NN is a hyperparameter that we must pick in order to get the best possible fit for the data set. Since one important way of judging the performance of any classification-procedure is to measure its "error rate" or misclassification probabilities, the lower the error rate, the better the procedure. Figure 4 shows the graphical representation of the error rate for Gaussian, Epanechnikov and tri-cube kernels according to the k values.

When k is small, we are restraining the region of a given prediction and forcing our classifier to be "more blind" to the overall distribution. In the case of geo-statistical interpolation, a small value of k produces dispersed patterns. In other words, a small value for k provides the most "flexible" fit, which will have low bias but high variance. On the contrary, a higher k averages more voters in each prediction and hence is more resilient to outliers: the interpolation map will have smoother decision boundaries which means lower variance but increased bias. In our work, we used two values of k depending on the dataset. For the manual labelled data, it was found that the convenient $k = 10$ which corresponded to a minimum error rate. In the case of automated labelling data, $k = 20$.

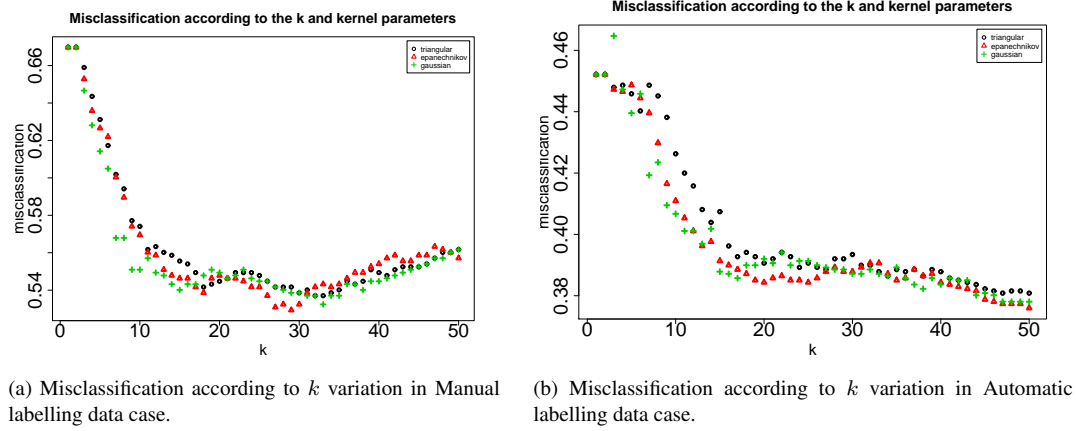


Figure 4: Error Rate for Gaussian, Epanechnikov and tri-cube kernels according to the k values.

4.2 Manual data labelling

Using Stein's terms (Stein 1999), "*interpolation means the predictions at locations that are "surrounded" by available observations or, alternatively, are not near or beyond the boundaries of the region in which there are observations*". Spatial distributions of the social strike phenomena can be approximated by functions depending on location in a multi-dimensional space. The interpolation in this case is the process of using points with known values (tweets location and sentiment polarity) to estimate values (sentiment polarity) at other unknown points location.

The social interpolation of Twitter data aims (in our case) to predict the value of a property for strike event at a location by using values of the same property sampled at scattered neighbouring points (Journel & Huijbregts 1978). This allows the estimation of the sentiment polarities at all points or nodes of a regular grid superimposed on a field of study on the basis of discrete points (scattered posted tweets).

The geographic points of the manually labelled data are plotted in Figure 5. The corresponding interpolation is presented in the Figure 6. The first thing we would like to point out is that the geographical zone associated with tweets that have no relation with the phenomenon of strike dominates the surface of France. This spatial coverage can be considered as noise for social network analysis. In addition, we would note that the colour gradient corresponds to the intensity of the tweets in relation to the polarity of feelings/opinions conveyed by the text. In particular, the adopted weighted scheme provides the probability estimation of each label in specific geographic area.

The opposition to the railway workers' strike is highlighted in the north-western part of France (Brittany). Also, the neutral opinions are expressed more in the north-east of France (Ardennes and Picardie). Opinions that support the strike are expressed in the extreme south of France towards the department of Bouches du Rhône.

We are interested on the analysis of the interpolation without including the labels which do not give us information on the strike. Tweets deemed as off-topic were

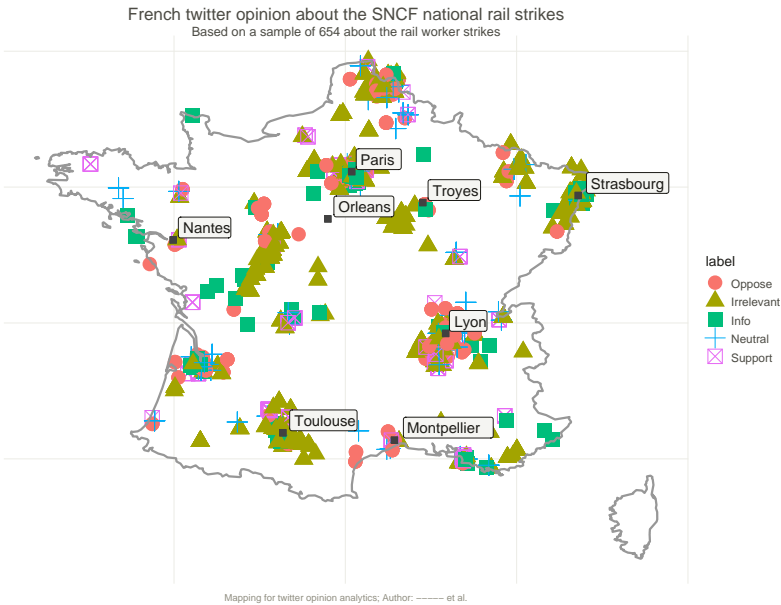


Figure 5: The map of Twitter data about the SNCF strike according to the manual labelling data.

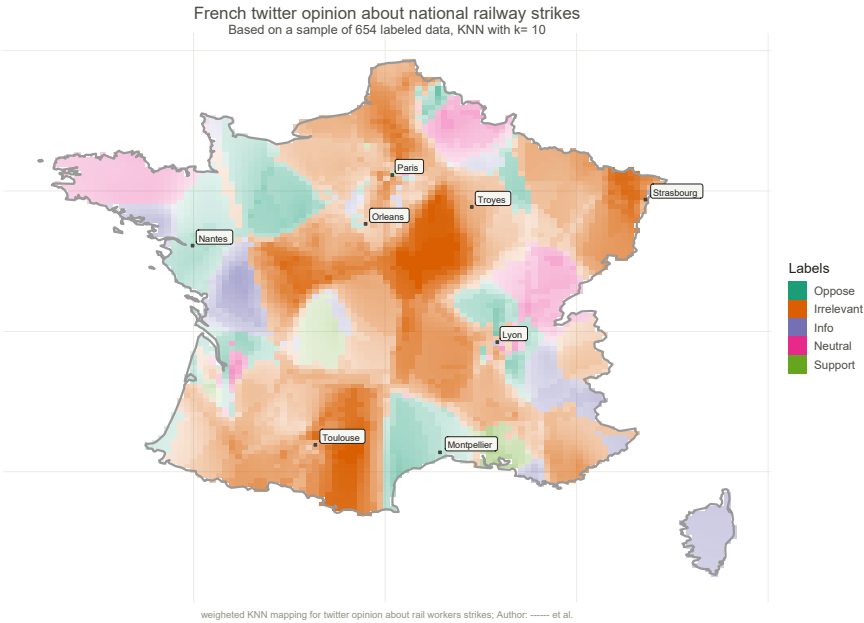


Figure 6: The weighted kernel k -NN interpolation over France for manual labelling data.

therefore eliminated in the analysis. This procedure allows us to highlight only the spatial characteristics related to the strike.

Figure 7 shows the data points for two kinds of categories: tweets opposing and supporting the railworkers' strike. The corresponding interpolation using a Weighted wk -NN model is presented in Figure 8.

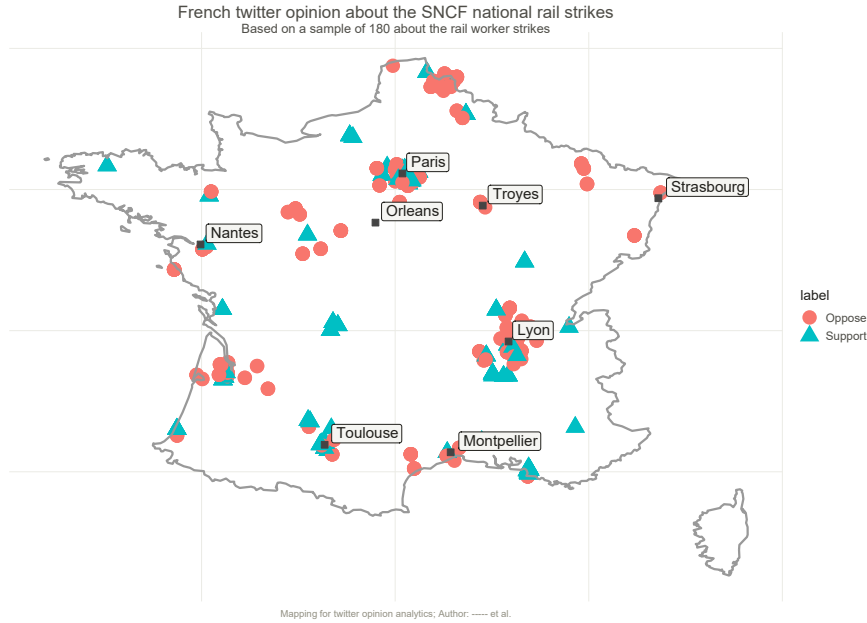


Figure 7: The map of Twitter data about the SNCF strike according to the manual labelling data .

The tweets clearly opposing the strike are expressed strongly over a specific geographical area in the north-east: the Nord-pas-de-Calais, Ardennes, Lorraine and Alsace regions. Although the western part is dominated by a feeling of opposition to the strike, there is nevertheless weaker expression: the probability gradient is clear over the whole region.

Tweets expressing a favourable opinion to the strike were observed especially in the sector from the France Centre to the South West (Toulouse). Haute-Normandie is represented by tweets expressing support for railway workers. It is also the case for the extreme South-East sector. It is very difficult to explain or interpret these results because the sample size in this case is very small. The suppression of off-topic tweets drastically reduced the amount of data processed.

4.3 Automatic data labelling

In order to obtain a more realistic picture and overcome the problem of sample size, automatic labelling was carried out on 4555 tweets. This is in order to have a fairly reasonable representativeness on the phenomena that provoked by the railwaymen strike. Only tweets expressing opposition or support for the strike were selected for this analysis.

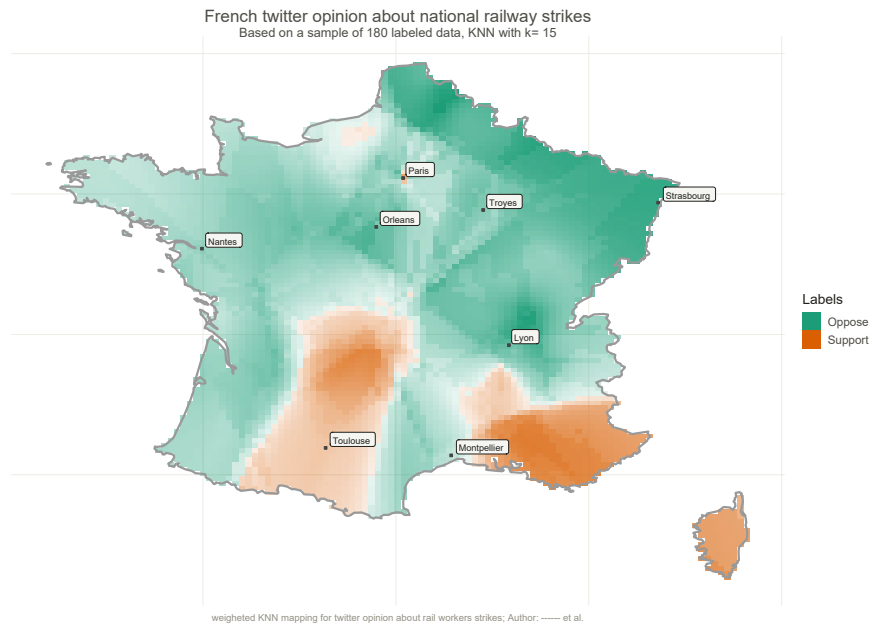


Figure 8: The map of Twitter data about the SNCF strike according to the manual labelling data.

The mapping of the automatically labelled data is shown in Figure 9. The corresponding spatial interpolation associated with these scattered points is given in Figure 10.

Comparing the two maps, Fig.8 and Fig.10, there is a slight difference especially in the North-Eastern part of France.

More generally, the opinion expressing a clear opposition to the strike dominates the entire territory of France. This is probably due to the cumulative problems of public transport in France. Users generally express their daily distress at the diminution in the quality of public services.

It should be noted that an opinion that opposes the strike does not necessarily mean support for the reform proposed by the government. It turns out that users only express their opposition to the strike because it severely affects their daily lives. Another aspect might be worth commenting on in this situation: there are transport users who support the movement of the strikers but are in disagreement with the protest mode. Several tweets point out the political implications of the reforms being pushed by the government, not just for the railways but for everyone.

5 Spatiotemporal trains regularity

In order to highlight the relationship between the services provided by SNCF and the perception of users towards these services, we chose to examine the regularity of Regional Express Transport (TER) trains. We expect that train punctuality is one of the most important factors that provides insights about the connection with the land and strikes. To do this,

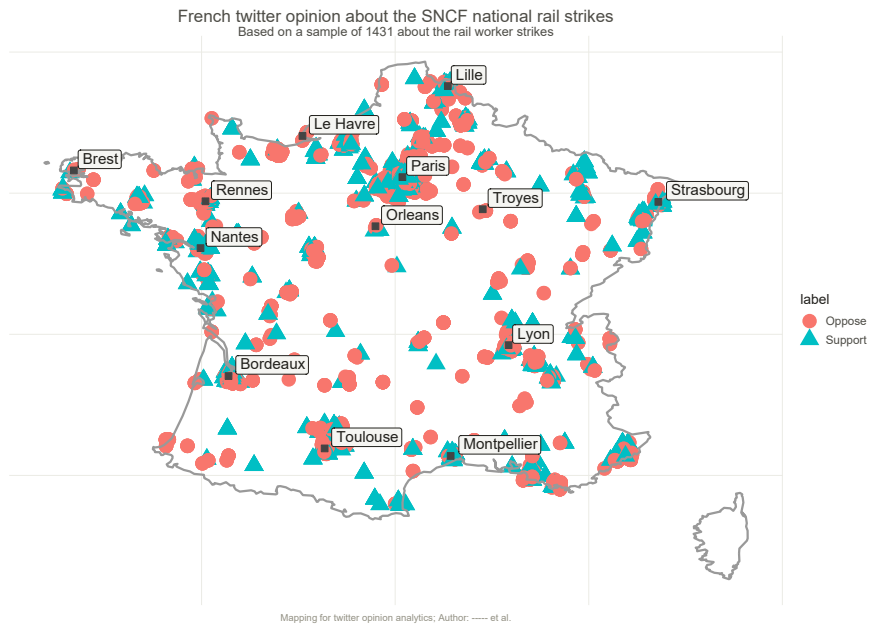


Figure 9: The map of Twitter data about the SNCF strike according to the automatic labelling data.

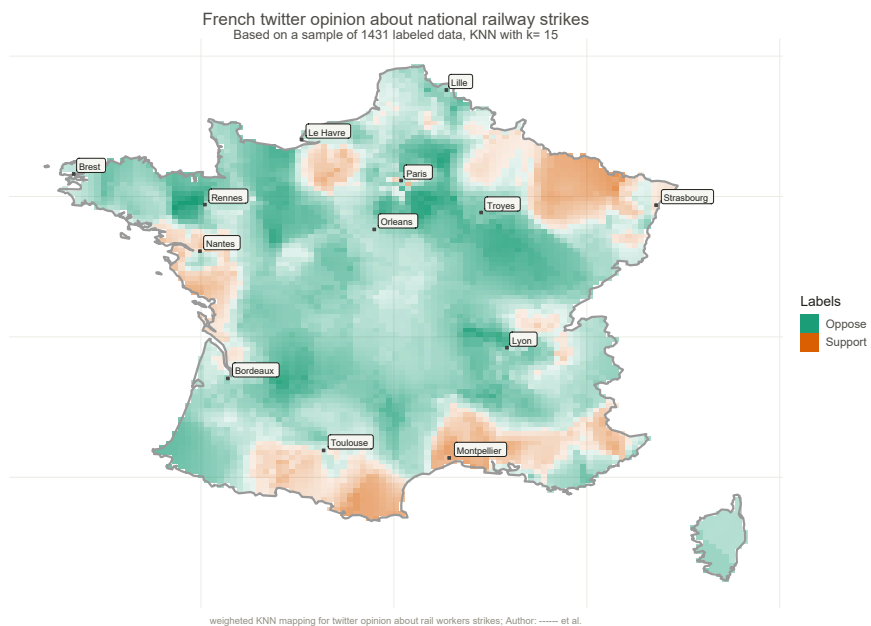


Figure 10: The map of Twitter data about the SNCF strike according to the automatic labelling data.

we retrieved the regularity and punctuality of regional trains data (TER Regional Express Transport) from **Open Data SNCF** platform.

Figure 11 shows two indicators of punctuality: number of concealed train and regularity. A cancelled train is a train whose programming was known to travellers and whose circulation was suppressed, even partially, without having been announced early enough. For simplification, a partial cancellation (on a part of the course) is assimilated to a total cancellation. The regularity is calculated when the train arrives at the last station of its route (terminus). This method of calculation, which does not include the intermediate stations, proposes the accumulated delay on the whole of a route. The indicator used is the “five-minute regularity”: a train is considered late if it arrives five minutes after its scheduled time. The proposed data are not detailed by TER line but aggregated for all TERs of a Region. Delays are those actually perceived by travellers and are therefore not relieved of any contractual neutralizations for exceptional external reasons.

All travellers are not equal in relation to the regularity of regional trains (TER). The analysis of SNCF data clearly draws two France: the North and the South, where the TER often arrive much more late and this when they are not cancelled at the last minute. Thus, three regions are distinguished by the regularity of their regional trains: Brittany (95.2% trains per hour), the Grand Est region (94.9%) and Normandy (93.78%). The Hauts de France, the Center-Val-de-Loire, the Pays de la Loire and Burgundy-Franche-Comté are a notch below, with a regularity rate close to 90%. The regularity falls to 88-87% for the new Aquitaine, Occitanie and the Auvergne-Rhône-Alpes region. Finally, the Provence-Alpes Côte d’Azur comes well behind with a regularity rate of 84.5%.

Provence-Alpes Côte d’Azur (PACA) is also the region with the most TER canceled (3.4%) with 520 canceled trains over the period of 2013 and 2017. On this second statistic, although always present, the North-South gap is more contrasted. After the PACA region, Occitania (2.28%), the new Aquitaine (1.98%), but also the Hauts-de-France (2.25%), are the regions with the most TER cancelled. In contrast to Brittany (0.9%). How to explain these differences? according to **Open Data SNCF**, social movements and other strikes are mentioned as minor causes of these disturbances, contrary to factors such as “work” (112 occurrences), “bad weather” (130 occurrences), “external” causes (154 occurrences) and other “material” concerns (162 occurrences).

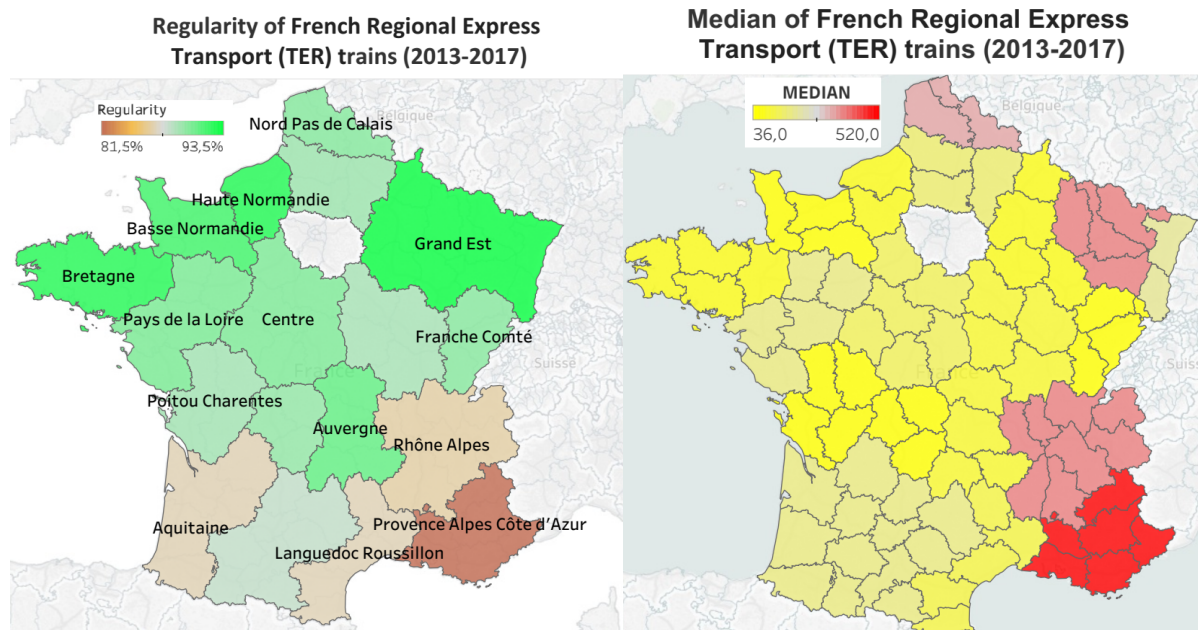
In terms of the temporal variability (Figure 12), things have not improved or deteriorated since 2013, the beginning of the collection of these regional data by the SNCF. The regularity rate and the cancellation rate do not vary much from one year to another. National averages remain stable, at around 91% of trains per hour and 2% of trains cancelled. However, if we distinguish the months of the year, the months of May and November clearly appear as critical moments.

For the regularity of the traffic, it is the month of November which stands out clearly. Among the reasons given by the regions, “wheel slippage” and “skidding” are multiplying, as well as “lack of grip” on railway tracks.

6 Discussion and conclusion

With nearly 29,000 km of operating lines and nearly 3,000 stations, based on these two criteria, the French rail network ranks second in Europe, behind Germany, (*cf* Figure 13a). France also has the second longest high speed network in Europe, behind that of Spain.

Figure 11: Regularity and median of canceled trains. Data of Regional Express Transport (TER) between 2013 and 2017 from **Open Data SNCF**.



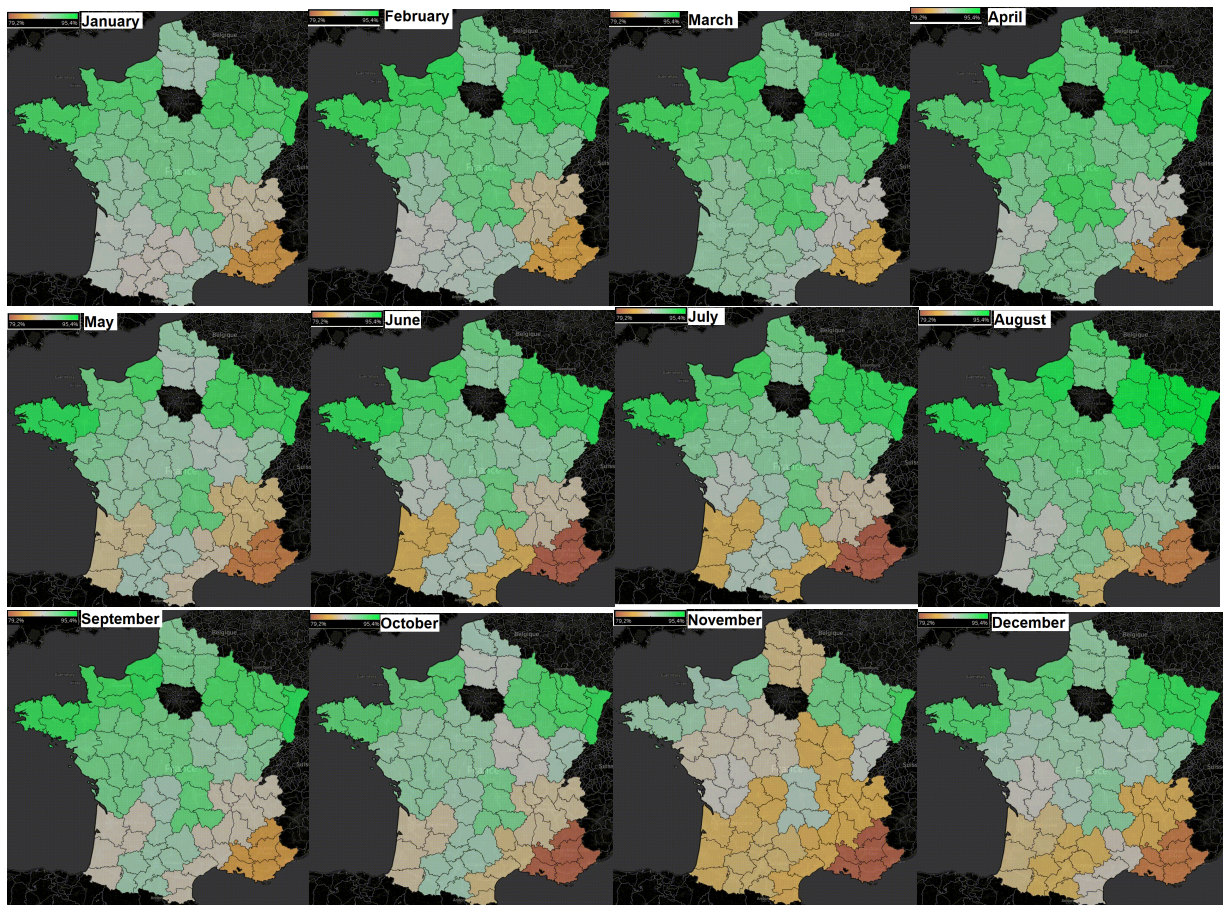
The French railways have been confronted for thirty years with the multiplication of their institutional interlocutors. A long period of direct dialogue with the State for the entire rail network has now been broken down into three types of services: national, regional and suburban, with very different organizing authorities and financing methods. Although the rail networks systems converge on Paris, the Ile-de-France region needs to be studied separately: French railway is highly polarized by Paris.

The disparity of opinion about the SNCF "reform" according to geographical regions may be due to the regional inequalities, for example territorial disparities of exposure to unemployment. That is, the so-called "spatial mismatch effect" according to which the physical distance between place of residence and place of work is a factor of exposure to unemployment and income inequality (*cf* Figure 13b). People with lower income tend to suffer more from restricted transport (poor or non-existent access) options (Lucas et al. 2016). This is central to restrictions on access to jobs, education and health facilities, social networks, etc. As showed by Lucas et al. transport-related exclusion is a combination of socio-economic barriers preventing marginalized groups from using public transportation and the absence of affordable, safe, and accessible public transportation (Lucas 2012).

The lower income socio-economical category have often lower quality transport services available to them and travel under worse conditions (ITF 2017). The feeling expressed by those people about the SNCF strike and transport is a symptom of social inequality rather than an opposition to strikers' demands.

The current research attempts to effectively utilize social media to identify the spatial patterns of the biggest strikes in France in 2018. We have outlined the accuracy of weighted method interpolation using Gaussian kernel. The spatial patterns: social polarity opinion and feelings were highlighted using *k*-nearest neighbours. The wealth of the geographic

Figure 12: Spatio-Temporal variation of the Regional Express Transport regularity.



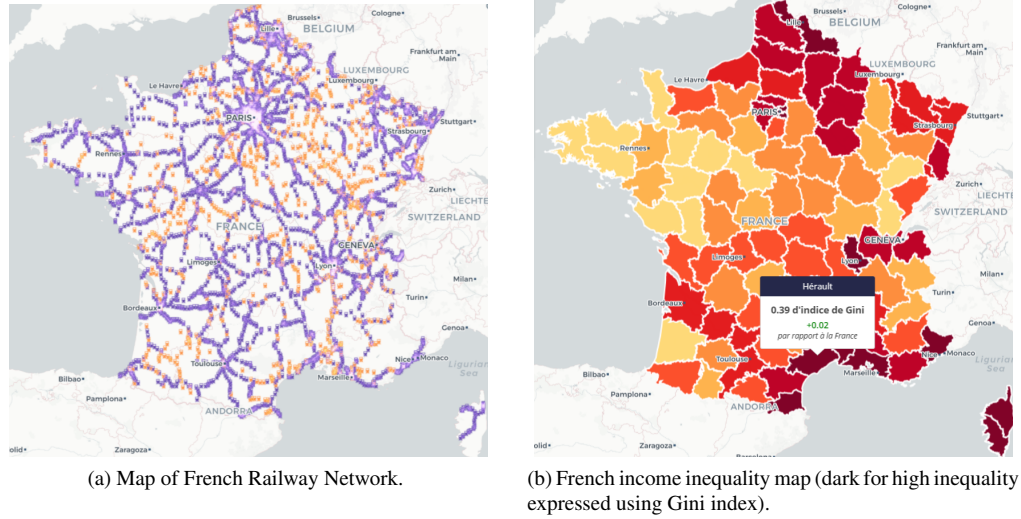


Figure 13: French income inequality and railway network

regions is correlated with the rail *irrigation* and the proximity of railway stations to large population centers. Bouf et al. showed that between 1990 and 2007, the GDP of regions in France served by the High-Speed Rail (TGV) increased by 42.1%, against 29.8% for regions without TGV connexion (Bouf & Desmaris 2014). These data is compatible with a "polarization effect" which favours the concentration of economic and demographic growth in the regions crossed by the TGV. The interpretation of different views expressed across France regions must include the relationships between daily travel behaviour and spatial, socio-economic characteristics, which needs exogenous factors.

The main contribution of this study is to tackle the social events problem as a geo-statistical learning for the social networks geographical issue. The developed geo-statistical approach makes it possible to quantify the uncertainty associated with the map and define the zones subsampled. In case of automatic labelling, we addressed data that the performance of the classifiers degraded from event data when out-of-event training samples were added to training samples.

The approach does have some bias, since the Twitter user-base is not representative of the population in general and the discontinuity of data points do not cover the entire space homogeneously.

We would like to emphasize that there is no single method that can be applied to all situations to interpret results obtained from spatial interpolation. In the remote areas of the observation points, it is noted that any attempt at characterization is difficult, but an approximative estimation is possible using the weighted kernel approach. Thus, the addition of ancillary data, which could take into account the external drift, should make it possible to improve the accuracy of the results further, and thus the ability to interpret them.

Research around social events polarity by analysing spatial patterns should extend more widely beyond the methodology introduced in this paper. Our work falls under events spatial interpolation perspective alone, and therefore limited in its description of all social events behaviours. Identifying strikes events and their relevant tweets is non-trivial. As Twitter is arguably the most popular microblog site for user to post and share, most tweets are about

586 daily routines and personal interests, while only a small proportion of tweets underlying
 587 the Twitter torrent is event related. For these “*traditional news*” dataset, event detection
 588 is well studied under Topic Detection and Tracking (TDT) in the information retrieval
 589 community (Allan et al. 1998, Yang et al. 1999). These work focuses on evolutionary
 590 clustering of streaming news articles. Nevertheless, identifying spatial events on Twitter
 591 stream is more challenging, since we need to unearth event related tweets from huge tweet
 592 flow (DIAO 2015). Weng and Lee proposed a method that first characterizes temporal
 593 patterns of individual words using wavelets and then groups them into events Weng & Lee
 594 (2011). For possible improvements of our study, we can complete the spatial characterization
 595 of strikes events by its temporal evolution. In other respects, there have been quite a few
 596 studies have explored supervised approaches to distinguishing between messages about
 597 real-world events and non-event messages for Twitter stream analysis (Becker et al. 2011).
 598 Our event filter in the preprocessing part can be inspired by these works as well as by
 599 topic modeling on Twitter (Hong & Davison 2010, Yang et al. 2014) and Natural Language
 600 Generation Saggion & Lapalme (2002), Jing & McKeown (2000).

601 Further improvements to the proposed approach could be obtained with the following:

- 602 • Comparison with other possible hybridizations, such as including economic,
 603 demographic and social parameters;
- 604 • An enhanced statistical feature with temporal variability analysis.

605 In addition, we have identified two main challenges: how enrich the training data
 606 sample and non-trivial human-labelling tasks, as our medium-term goals for potential future
 607 improvements.

608 References

- 609 Allan, J., Papka, R. & Lavrenko, V. (1998), On-line new event detection and tracking., *in*
 610 ‘Sigr’, Vol. 98, Citeseer, pp. 37–45.
- 611 Bailey, T. & Jain, A. (1978), ‘A note on distance-weighted k -nearest neighbor rules’, *IEEE*
 612 *Transactions on Systems, Man, and Cybernetics* (4), 311–313.
- 613 Becker, H., Naaman, M. & Gravano, L. (2011), Beyond trending topics: Real-world event
 614 identification on twitter, *in* ‘Fifth international AAAI conference on weblogs and social
 615 media’.
- 616 Bicego, M. & Loog, M. (2016), Weighted k-nearest neighbor revisited, *in* ‘Pattern
 617 Recognition (ICPR), 2016 23rd International Conference on’, IEEE, pp. 1642–1647.
- 618 Bollen, J., Mao, H. & Zeng, X. (2011), ‘Twitter mood predicts the stock market’, *Journal*
 619 *of computational science* 2(1), 1–8.
- 620 Bouf, D. & Desmaris, C. (2014), ‘Trains à grande vitesse et équité spatiale en france’.
- 621 Brzozowski, M. J. & Romero, D. M. (2011), Who should i follow? recommending people
 622 in directed social networks., *in* ‘ICWSM’.
- 623 Castells, M. (2015), *Networks of outrage and hope: Social movements in the Internet age*,
 624 John Wiley & Sons.

- 625 Courant, R. & Hilbert, D. (1962), *Methods of Mathematical Physics*, number vol. 1 in
626 ‘Methods of Mathematical Physics’, Interscience Publishers.
- 627 Cover, T. & Hart, P. (1967), ‘Nearest neighbor pattern classification’, *IEEE transactions on*
628 *information theory* **13**(1), 21–27.
- 629 DIAO, Q. (2015), Event identification and analysis on Twitter, PhD thesis.
- 630 Dudani, S. A. (1976), ‘The distance-weighted k-nearest-neighbor rule’, *IEEE Transactions*
631 *on Systems, Man, and Cybernetics* (4), 325–327.
- 632 Fix, E. & Hodges Jr, J. L. (1951), Discriminatory analysis-nonparametric discrimination:
633 consistency properties, Technical report, California Univ Berkeley.
- 634 Gaber, I. (2017), ‘Twitter: A useful tool for studying elections?’, *Convergence* **23**(6), 603–
635 626.
- 636 Hanspeter, K., Koopmans, R., Duyvendak, J. W. & Giugni, M. G. (2015), *New social*
637 *movements in Western Europe: A comparative analysis*, Routledge.
- 638 Hechenbichler, K. & Schliep, K. (2004), Weighted k-nearest-neighbor techniques and
639 ordinal classification, Technical report, University Munich.
- 640 Hong, L. & Davison, B. D. (2010), Empirical study of topic modeling in twitter, in
641 ‘Proceedings of the first workshop on social media analytics’, acm, pp. 80–88.
- 642 Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F. & Rahman, R. M. (2015), ‘Localized
643 twitter opinion mining using sentiment analysis’, *Decision Analytics* **2**(1), 8.
- 644 ITF (2017), Income inequality, social inclusion and mobility, Technical report, Organisation
645 for Economic Co-operation and Development, The International Transport Forum .
- 646 Jing, H. & McKeown, K. R. (2000), Cut and paste based text summarization, in ‘Proceedings
647 of the 1st North American chapter of the Association for Computational Linguistics
648 conference’, Association for Computational Linguistics, pp. 178–185.
- 649 Journel, A. G. & Huijbregts, C. J. (1978), *Mining geostatistics*, Vol. 600, Academic press
650 London.
- 651 Lucas, K. (2012), ‘Transport and social exclusion: Where are we now?’, *Transport policy*
652 **20**, 105–113.
- 653 Lucas, K., Mattioli, G., Verlinghieri, E. & Guzman, A. (2016), Transport poverty and
654 its adverse social consequences, in ‘Proceedings of the institution of civil engineers-
655 transport’, Vol. 169, Thomas Telford (ICE Publishing), pp. 353–365.
- 656 MacLeod, J. E., Luk, A. & Titterington, D. M. (1987), ‘A re-examination of the distance-
657 weighted k-nearest neighbor classification rule’, *IEEE Transactions on Systems, Man,*
658 *and Cybernetics* **17**(4), 689–696.
- 659 Mercea, D. & Yilmaz, K. E. (2018), ‘Movement social learning on twitter: The case of the
660 people’s assembly’, *The Sociological Review* **66**(1), 20–40.

- 661 Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J.,
662 Cohen, J., Callegaro, M. & Harwood, P. (2014), 'Social media in public opinion research:
663 executive summary of the aapor task force on emerging technologies in public opinion
664 research', *Public Opinion Quarterly* **78**(4), 788–794.
- 665 Paroubek, P., Grouin, C., Bellot, P., Claveau, V., Eshkol-Taravella, I., Fraisse, A., Jackiewicz,
666 A., Karoui, J., Monceaux, L. & Torres-Moreno, J.-M. (2018), Deft2018 : recherche
667 d'information et analyse de sentiments dans des tweets concernant les transports en île
668 de france, in 'Actes de DEFT', Rennes, France.
- 669 Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. & Mozetič, I. (2015), 'The effects of
670 twitter sentiment on stock price returns', *PloS one* **10**(9), e0138441.
- 671 Rosen, A. & Ihara, I. (2017), 'Giving you more characters to express yourself', *Twitter Blog*
672 **26**.
- 673 Saggion, H. & Lapalme, G. (2002), 'Generating indicative-informative summaries with
674 sumum', *Computational linguistics* **28**(4), 497–526.
- 675 Statista (2018), 'Number of monthly active twitter users worldwide from 1st quarter 2010
676 to quarter 2018 (in millions)'. Accessed: 2018-08-31.
- 677 Stein, M. L. (1999), *Interpolation of spatial data: some theory for kriging*, Springer.
- 678 Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010), 'Predicting elections
679 with twitter: What 140 characters reveal about political sentiment.', *Icwsm* **10**(1), 178–
680 185.
- 681 Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2011), 'Election forecasts
682 with twitter: How 140 characters reflect the political landscape', *Social science computer
683 review* **29**(4), 402–418.
- 684 Twitter (2018), 'Tutorials filtering tweets by location', [https://developer.twitter.com/en/docs/tutorials/
685 filtering-tweets-by-location.html](https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.html). Accessed: 2018-10-10.
686 **URL:** <http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm>
687
- 688 Weng, J. & Lee, B.-S. (2011), Event detection in twitter, in 'Fifth international AAAI
689 conference on weblogs and social media'.
- 690 Yang, S.-H., Kolcz, A., Schlaikjer, A. & Gupta, P. (2014), Large-scale high-precision topic
691 modeling on twitter, in 'Proceedings of the 20th ACM SIGKDD international conference
692 on Knowledge discovery and data mining', ACM, pp. 1907–1916.
- 693 Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T. & Liu, X. (1999),
694 'Learning approaches for detecting and tracking news events', *IEEE Intelligent Systems
695 and their Applications* **14**(4), 32–43.
- 696 Zhang, X., Fuehres, H. & Gloor, P. A. (2011), 'Predicting stock market indicators through
697 twitter "i hope it is not as bad as i fear"', *Procedia-Social and Behavioral Sciences* **26**, 55–
698 62.