CoViD-19, learning from the past: A wavelet and cross-correlation analysis of the epidemic dynamics looking to emergency calls and Twitter trends in Italian Lombardy region

Bruno Alessandro Rivieccio[1*¶], Alessandra Micheletti[2¶], Manuel Maffeo[3,4], Matteo Zignani[5], Alessandro Comunian[6], Federica Nicolussi[7], Silvia Salini[7], Giancarlo Manzi[7], Francesco Auxilia[3,8], Mauro Giudici[6], Giovanni Naldi[2], Sabrina Gaito[5], Silvana Castaldi[3,9&], Elia Biganzoli[10&]

[1] Department of Laboratory Medicine, Division of Anatomic Pathology, Niguarda Hospital, Milan, Italy

[2] Department of Environmental Science and Policy, University of Milan, Milan, Italy

[3] Department of Biomedical Sciences for Health, University of Milan, Milan, Italy

[4] Public Health Post Graduate School, University of Milan, Milan, Italy

[5] Department of Computer Science, University of Milan, Milan, Italy

[6] Department of Earth Sciences, University of Milan, Milan, Italy

[7] Department of Economics, Management and Quantitative Methods & Data Science Research Center, University of Milan, Milan, Italy

[8] ASST FBF-Sacco, Milan, Italy

[9] Fondazione IRCCS Ca' Granda Ospedale Maggiore, Milan, Italy

[10] Department of Clinical Sciences and Community Health & Data Science Research Center, University of Milan, Milan, Italy

[*] Corresponding author

E-mail: brunoalessandro.rivieccio@ospedaleniguarda.it

[¶] These authors contributed equally to this work.

[&] These senior authors also contributed equally to this work.

Abstract

The first case of Coronavirus Disease 2019 in Italy was detected on February the 20th in Lombardy region. Since that date, Lombardy has been the most affected Italian region by the epidemic, and its healthcare system underwent a severe crisis during the outbreak. From a public health point of view, therefore, it is fundamental to provide healthcare services with tools that can reveal a possible new epidemic burden with a certain time anticipation, which is the main aim of the present study. Moreover, the sequence of law decrees to face the epidemic and the large amount of news generated in the population feelings of anxiety and suspicion. Considering this whole complex context, it is easily understandable how people "overcrowded" social media with messages dealing with the pandemic, and emergency numbers were overwhelmed by the calls. Thus, in order to find potential predictors of a possible second epidemic wave, we analyzed data both from Twitter and from emergency services comparing them to the daily infected time series at a regional level. Since our principal goal is to forecast a possible new ascending phase of the epidemic, we performed a wavelet analysis in the time-frequency plane, to finely discriminate over time the anticipation capability of the considered potential predictors. In addition, a cross-correlation analysis has been performed to find a synthetic indicator of the time delay between the predictor and the infected time series. Our results show that Twitter data are more related to social and political dynamics, while the emergency calls trends can be further evaluated as a powerful tool to potentially forecast a new burden. Since we analyzed aggregated regional data, and taking into account also the huge geographical heterogeneity of the epidemic spread, a future perspective would be to conduct the same analysis on a more local basis.

Introduction

On February the 20th the first Italian case of Coronavirus Disease 2019 (CoViD-19) due to secondary transmission outside China was identified in Codogno, Lombardy region [1]. In the following days the number of cases started to rise not only in Lombardy but also in other Italian regions, although Lombardy remained and is still the most affected region in Italy [2]. At the time of writing (October the 4th), 325,329 cases have been identified in Italy, out of which 108,065 in Lombardy region [3].

The progressive decrease of CoViD-19 cases should not let our guard down, indeed it is clear that, since the beginning of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic, public health has suffered from the absence of a proper preparedness plan to face an episode which was unexpected and unpredictable and has heavily impacted on the territorial and hospital healthcare services. Planning has a fundamental role nowadays but it can be adequate only if the next possible pandemic peak can be effectively foreseen by means of a predictive tool which accounts for all the available signals. In order to do so, it is of paramount importance to learn from what happened during the first peak to be prepared for the potential next one.

The SARS-CoV-2 outbreak in Italy has been characterized by a massive spread of news coming from both official and unofficial sources leading to what has been defined as infodemia, an over-abundance of information – some accurate and some not – that has made hard for people to find trustworthy sources and reliable guidance needed [4].

Infodemia on SARS-CoV-2 created the perfect field to build suspicion in the population, which was scared and not prepared to face this outbreak. It is understandable how the rapid increase of the cases number, the massive spread of news and the adoption of laws to face this outbreak led to a feeling of anxiety in the population, whose everyday life changed very quickly.

A way to assess the dynamic burden of social anxiety is a context analysis of major social networks activities over the internet. To this aim Twitter represents a possible ideal tool, because of the focused role of the tweets according to the more urgent needs of information and communication, rather than general aspects of social projection and debate as in the case of Facebook, which could provide slower responses for the fast individual and social context evolution dynamics [5].

Taking into account this specific context, it is easy to understand why the 112 emergency number service in Lombardy region was suddenly overwhelmed by an enormous number of calls that rapidly overcame its capacity to cope and compromised the possibility to identify those patients who needed immediate medical assistance [6].

As pointed out by the Scientific Italian Society for Medical Emergency (SIEMS), the number of calls to 112 for the Milan province was 5,086 on February the 16th, before the outbreak, and rapidly increased to 6,798 on February the 21st and to 10,657 on February the 22nd [7].

The emergency service in Lombardy region is organized through three first-level PSAPs (public-safety answering points), called CUR-NUE (Unique answering operating room / point – European emergency number), which forward the call to the most appropriate service, i.e. Police, Fire Department or Medical emergency rescue service. So, after the first assessment, calls requiring medical assistance are sent to one of the four second-level PSAPs called SOREU (Regional Operating Rooms for Medical Emergency and Urgency), depending on the geographical area the call is coming from in order to evaluate the patient and decide the most appropriate intervention.

Lombardy region, from an administrative point of view, is made up of twelve provinces, and the management of medical emergency calls occur in four operating rooms which receive these calls from different sets of provinces, namely:

- SOREU delle Alpi (SRA) from the provinces of Bergamo, Brescia and Sondrio;
- SOREU dei Laghi (SRL) from the provinces of Como, Lecco and Varese;
- SOREU Metropolitana (SRM) from the provinces of Milan and Monza-Brianza;

- SOREU della Pianura (SRP) from the provinces of Cremona, Lodi, Mantova and Pavia.

Some of the non-urgent calls received by the four SOREU do not need, though, an ambulance dispatch, so not all the calls result in a medical rescue mission: in the latter cases the patients are recommended by SOREU medical technicians to consult other medical services such as general practitioners. Moreover, during the epidemic, according to specific internal procedures, SOREU medical technicians answering to the emergency calls, in the case of signs and/or symptoms evocative of CoViD-19 but not life threatening, advised patients to wait for a recall by a public health medical doctor: after this re-evaluation call, it was up to these medical doctors the final decision on the management of the case (ambulance dispatch or home quarantine).

To reduce the burden of calls to the emergency number which occurred during the first days of the outbreak, it was necessary to redirect non-urgent calls, especially those asking for information, to other services. According to European Emergency Number Association guidelines [8], Lombardy region created a regional toll-free number for CoViD-19, the first one in Italy. Other Italian regions created their own one in the following weeks, as well as other European countries like Spain, Germany, Croatia, which were facing similar issues [9].

The 24/24-hour toll-free number was settled on February the 23rd by AREU (Regional Emergency Service Agency) and, although it helped to funnel non-urgent calls, it was not enough because of the huge number of calls: for example on the second day it received more than 400,000 calls.

Calls to the emergency services could be an important and helpful indicator of the spread of the infection among the population, taking into account the possibility to analyze data regarding the municipality from which the calls originated and the motivations that induced people to ask for fast medical support. Statistical models could be used to assess the association of these data with new cases of CoViD-19 in order to predict new epidemic hotspots on a municipal scale, or with a smaller spatial scale for big cities.

In addition to usual public health indicators, social media data may also be used as probes of the people behaviour according to the recent trends of digital epidemiology. As mobile technology continues to evolve and proliferate, social media are expected to occupy an increasingly prominent role in the field of infectious diseases [10-12] .

A recent systematic review concluded that the inclusion of online data in surveillance systems has improved the disease prediction ability over traditional syndromic surveillance systems and Twitter was the most common social network analyzed for this aim [13].

Despite some limitations and concerns, a better understanding of the behavioural change induced by social media can strengthen mathematical modelling efforts and assist in the development of public policy so as to make the best use of this increasingly ubiquitous resource in controlling the spread of disease [11].

Aim of the study is to understand the correlation between the users calls to the emergency services and the spread of the infection in the population during the first peak of the CoViD-19 outbreak in Lombardy region of Italy, the first world hotspot after the Chinese raise in Wuhan. Furthermore, the joint analysis with Twitter trends related to emergency was performed to better understand the most important population concerns according to the infection dynamics. Overall, the joint active monitoring of the communication dynamics over emergency calls and social networks like Twitter could provide an integrated means for the adaptive management of information delivery as well as the optimization of the rescue logistic and finally it could provide relevant anticipation on the outbreak. These aspects appear of critical importance for CoViD-19 surveillance, and for the preparedness of emergency and strategic plans [14].

Materials and methods

Data

In the present work, we analyzed the following time series:

- PSAP-II SOREU-118 daily incoming calls from 2020.02.18 to 2020.03.30 [15];
- PSAP-I CUR NUE-112 daily incoming calls from 2020.02.18 to 2020.03.30 [16];
- toll-free number daily incoming calls from 2020.02.23 to 2020.03.30 [16];
- daily Twitter data (tweets, replies, likes, retweets) from 2020.02.18 to 2020.06.29 [17];
- daily infected from 2020.02.24 (the first day since which Italian Department for Civil Defense has provided data) to 2020.06.29 [3].

Data about SOREU-118, NUE-112 and toll-free number daily incoming calls were only available at a regional level. The very first days of the NUE and toll-free number time series have been discarded due to the very intense population panic reaction which reflected into a very huge amount of calls (whose peak was even higher than the following, new cases-related one). Particularly, in the case of NUE they were inappropriate non-urgent calls (most of all for information need), so they were not forwarded to the corresponding SOREU: indeed, in the SOREU time series we do not observe any peak in the very first days. Moreover, this choice is justified if we consider that – in case of a new epidemic burden – there would not be such a powerful reaction, so to the aim of predictability we can take into account just the subsequent new increase in the calls to NUE and toll-free number, which is more related to the CoViD-19 dynamics. Twitter data, instead, differently from the emergency calls, were not geolocalized. Finally, daily new cases have been collected at the province level and then aggregated at the regional level.

Twitter data analysis

The monitoring of the communication dynamics on online social media has been conducted on Twitter [18]. Specifically, the Twitter Search API (Application Programming Interface) was used to collect all the tweets in Italian language containing the keywords “112” or “118” in the body text. The data span the period from 2020.02.18 to 2020.06.29. In addition, the text of the tweets was further filtered to identify the most common keywords related to the emergency, bringing to the manual identification of 290 keywords. To identify them, a POS (Part of Speech)-tagger was run on each tweet, keeping the lemma of nouns and verbs. Then, all the terms which occurred in the tweet corpus less than five times were discarded. Finally, all the tweets that did not contain any keyword in the text were excluded from the Twitter dataset, leading to 5,989 statuses used for the purpose of this paper. Since both the timestamp and the number of likes, retweets and replies at the moment of the data collection were available for each tweet, it was possible to reconstruct four time series related to the dynamics of the emergency calls on Twitter: 1) the production of new statuses per day, 2) the number of retweets of new statuses per day, 3) the number of likes per day, and 4) the number of replies per day. The latter is a first indicator of the level of discussions triggered by the tweets, while the second one and the third one represent an early approximation of the endorsement to the content of the tweet.

Wavelet analysis

Wavelets represent a powerful tool to analyze localized variations of non-stationary power at many different frequencies in time series: the decomposition of the signal information in the time-frequency domain, indeed, allows to bring out variability and its changes over time [19-22]. Particularly, wavelets can “capture” and detect in the time-frequency (scale) plane long-period (low frequency) trends (properly called trends or backgrounds), and short-period (high frequency) trends (called anomalies or discontinuities). Anomalies, despite their limited spatio-temporal location,

possess a huge amount of information content, thus it is of primary importance to reveal them adequately. Among the others, the wavelet transform has relevant features such as a good capability of time-frequency localization (useful to analyze signals changing over time), and offers the possibility of a multi-resolution representation over different scales [19]. Wavelets transforms, differently from other series expansions, are thus a suitable tool to identify both a trend in a nonparametric form, keeping significant local peaks, and variations from the trend [23-24].

In order to identify the trends, all the time series were first smoothed using a moving average linear filter of a 7-days amplitude, decomposed through the wavelet transforms and then normalized to their maximum values. Tweets-dependent data (daily number of replies, likes and retweets) were also previously normalized to the correspondent number of daily tweets.

Details on the adopted continuous wavelet transform (CWT) and related measures are reported in the Supplementary material. Briefly, first of all, the CWT of all the signal was computed, together with the magnitude of the wavelet transform (using the modulus of the complex values). After that, for each pair of time series, the following quantities were calculated: the magnitude-squared wavelet coherence (MSWC), the wavelet cross-spectrum (WCS), the cone of influence, the phase coherence relationship (using the argument of the complex values), the time delay between the two signals (using the phase lag values). Each value of scale was converted to the equivalent Fourier value of frequency, and thus to the correspondent period. However, note that the relationship between scale and frequency is only an approximation, since there is not a precise correspondence between the two: among others, Meyers and coll. (1993) proposed a method for the conversion from scales to “pseudo-frequencies” [25].

Wavelet coherence and cross-spectrum analysis provides a detailed both time- and frequency-localized information on the phase lag and thus on the time delay between the compared signals, thanks to the decomposition through CWT, but lacks a global view of the trends and their relative shifts over time. Thus, in order to give a synthetic and unique indicator of the similarity

over time between the signals, we also performed a time domain analysis estimating their cross-correlation sequence.

Time lag estimation through cross-correlation analysis

In signal processing, another way to determine the similarity of two discrete time sequences is the cross-correlation. Indeed, since cross-correlation measures the relation between a vector x and time-shifted (lagged) copies of another vector y as a function of the lag itself, it allows also a time-delay analysis between the signals. Moreover, by revealing their relative displacement in time, the lag at which there is the maximum correlation can be considered as the time shift necessary to align the series x and y by sliding y backward (negative lag, in case y has a delay compared to x) or forward (positive lag, in case y displays an anticipation respect to x). To this aim, we computed the cross-correlation function between the time series, at different lags. The lag with the maximum cross-correlation value was thus identified as the "typical time delay" leading the lagged signal.

A 90% confidence interval was computed both for the cross-correlation values and for the lag corresponding to the cross-correlation function peak, i.e. for the days of delay. We used two different methods to calculate the interval estimation in order to compare parametric and non-parametric estimates: 1) a Fisher's z statistics through the Fisher's z-transformation [26]; 2) a Monte Carlo method through a surrogate time series with the same auto-correlation of the original one, for 1,000 simulations [27]. Details and notes on these two confidence interval computation methods are provided in the Supplementary material.

Cross-correlation could also be estimated using wavelets, specifically through a maximal overlap discrete wavelet transform (MODWT): it would have been our intention to complete the analysis with wavelets, but since emergency calls time series are made up of few samples, there are not enough non-boundary coefficients, even at the first level, to compute the wavelet cross-

correlation function for a sufficient number of lags (see explanation about the edge effects in the Supplementary material).

All the previous reported analyses were performed using MatLab R2020a, The MathWorks Inc.

Results

Wavelet analysis

As already pointed out before, the following daily regional aggregated data were considered:

- toll-free number incoming calls,
- NUE-112 incoming calls,
- SOREU-118 incoming calls.

In addition, we compared also Twitter data to the number of regional daily infected patients.

The WCS and the MSWC were calculated for each of these time series in relation to the data of regional daily infected. Indeed, both wavelet cross-power spectrum and coherence, through the CWT, can show areas in the time-frequency space where two signals share common harmonic components. In particular, the focus will be on the areas for which coherence is higher than 0.5 (indicated by the arrows, see Figs 1-7), since for lower values of coherence the phase lag is not reliable.

In the following figures (Figs 1-7), the time courses of the smoothed and normalized series are displayed on the left, whereas the WCS/MSWC representation is on the right. In the time courses charts, the daily infected curve is in red, while the potential predictor curve is in blue. In the WCS/MSWC charts, the *x* axis represents time (days), the *y* axis (logarithmic scale) represents scale

(which has been converted to the equivalent Fourier frequency, cycles/day), and the color scale represents the MSWC. The cone of influence, where edge effects should be considered, is shown as a white dashed line. For areas where the coherence exceeds 0.5, the charts display arrows to show the phase lag between the two signals. The arrows, which do not represent vectors since the length is not proportional to the intensity, are spaced in time and scale. The direction of the arrows designates the relative phase on the unit circle: a rightward-pointing arrow indicates in-phase coherence relationship of the two signals ($\Delta\varphi = 0$); a leftward-pointing arrow indicates anti-phase coherence relationship ($\Delta\varphi = \pi$). The corresponding lag in time depends on the duration of the cycle (period).

Fig 1. Regional toll-free number daily incoming calls vs. daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (toll-free number calls in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 2. Regional NUE daily incoming calls vs. daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (NUE calls in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 3. Regional SOREU daily incoming calls vs. daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (SOREU calls in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 4. Daily number of tweets vs. regional daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (tweets in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 5. Daily number of likes vs. regional daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (likes in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 6. Daily number of retweets vs. regional daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (retweets in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Fig 7. Daily number of replies vs. regional daily infected time courses and wavelet analysis. On the left, the smoothed (7-days amplitude moving average) and normalized time courses are displayed (replies in blue, daily infected in red); on the right, WCS and MSWC chart is shown (see text for explanation).

Looking at the time courses, it is evident a large anticipation (about two weeks) of the emergency calls trends with respect to the epidemic dynamics: if we consider the peaks of the curves, indeed, while the infected time series reaches its maximum at day 29, the toll-free number and SOREU calls peak occur at day 15 (time delay -14 days), and the NUE calls curve reaches the peak at day 16 (time delay -13 days). However, wavelet analysis does not reveal any strong coherence for all these pair of signals: specifically, among all the previously mentioned time series,

the only one for which wavelet analysis does not display any relevant coherence compared to daily new cases data is that of regional toll-free number incoming calls (Fig 1). Instead, NUE regional data and daily infected signals (Fig 2) display coherence over days from 18 to 22 at frequencies around 0.25 cycles/day, with a phase lag from -126.4° to -134.7°, corresponding to a time delay from 2.5 days to 2.6 days. Interestingly, days from 18 to 22 are confined between the two peaks, since the NUE calls curve reaches the peak at day 16, while the infected curve reaches its maximum at day 29. Not surprisingly, wavelet cross-spectrum and coherence analysis between regional daily incoming calls to SOREU and infected people (Fig 3) shows an anomaly less limited over time and over scale (frequency band ranging from 0.25 cycles/day to 0.37 cycles/day, time interval from day 17 to day 28), with a phase shift between -72.1° and -111.2°, leading to a time delay of about 2.4 days.

Even if not geolocalized, we finally compared regional epidemic time series with Twitter data (Figs 4–7): just considering the time courses, it is evident that the best potential predictor is the tweets time series (peaks at day 22 and day 29, respectively for daily number of tweets and infected curve). Indeed, since replies, likes and retweets are variables dependent on the original tweets, they are delayed in time: likes and retweets reach the peak at day 28, just one day before new cases, loosing almost all the anticipation capability, while the maximum value of replies is achieved even later than the peak of the infected curve, at day 33. Wavelet analysis, focused on the predictability, confirms this consideration:

- in the case of daily tweets (Fig 4) it detects two relevant areas of high coherence in the time-frequency plane:
    - a trend at the lowest frequencies (from 0.02 cycles/day to 0.03 cycles/day), with the phase relationship shifting from in-phase coherence to a maximum lag of 31.8° (time delay 2.6 days),

- and a time-localized anomaly before the infected peak around the frequency of 0.25 cycles/day, with a phase-lag ranging from 73.0° (day 20, 0.37 cycles/day) to 116.3° (day 27, 0.23 cycles/day) and a subsequent time delay between 0.5 and 1.4 days;

- for the number of daily likes (Fig 5) just a trend localized in the first half of the observation period and at the lowest frequencies is captured, with a maximum phase lag of 65.3° at 0.06 cycles/day, corresponding to a time shift of 3.0 days;
- daily number of retweets (Fig 6) display a similar trend more localized in the frequencies domain (from 0.04 cycles/day to 0.08 cycles/day), with a phase shift always near 0° (in the range between -26.7° and 19.6°);
- finally, the replies time series (Fig 7) shows a background even more localized both in time and frequency, and an anomaly around the frequency of 0.14 cycles/day from day 19 to day 33, both with a coherence value around 0.6.

Twitter trends

The next figures (Figs 8-9) display raw data about Twitter trends (tweets, replies, retweets and likes).

Fig 8. Twitter trends (1). The trends of daily number of new statuses about emergency calls (cyan line) and of replies (blue line) they sparked are shown here. The vertical green dotted lines indicate the principal episodes related to the lockdown policies in Italy.

Fig 9. Twitter trends (2). The trends of daily number of retweets (green line) and of likes (red line) are shown here.

It is evident that, concerning the dynamics of communications and the context on social media, Twitter activity (Figs 8–9) is not so strictly related to the epidemic dynamics, since it is triggered most of all by social, political and chronicle news, which drive an emotional participation of the users. Indeed, the first increase in all these time series (tweets, replies, likes and retweets), from day -3 to day 1, precedes just the establishment of the red areas in Codogno and Vo' Euganeo, while the second peak of daily tweets (Fig 8) at day 20 is related to the death of an operator of SRA due to CoViD-19. Moreover, likes and retweets (Fig 9) trends look more aligned to the announcements about lockdown policies.

Cross-correlation and time delay analysis

While wavelet decomposition allows a detailed analysis in both the time and frequency domains, it lacks a unique, global indication of the shift over time between two signals: to this aim, we performed also a time delay analysis estimating the cross-correlation sequence for each pair of time series. This analysis has been conducted only for the three cases for which WCS/MSWC revealed a strong coherence in the ascending phase (before both the peaks or between them), which is the most important from a public health monitoring point of view. These time series are the following: (i) daily regional incoming calls to NUE-112 (Fig 10); (ii) daily regional incoming calls to SOREU-118 (Fig 11); (iii) daily number of new tweets (Fig 12). In the following figures, the maximum of each function is depicted in red, and the confidence limits for the peak lag, deduced by the 90% *z*-Fisher confidence bounds of the cross-correlation values, are reported below the figures, in the caption. Negative lags denote by how many days the time series of infected patients should be shifted backward over time to be "aligned" with the predictor.

Fig 10. Cross-correlation sequence estimate between NUE regional incoming calls and daily infected. The blue lines represent the 90% confidence interval (CI) limits computed through a *z*-

transformation. The maximum value of the cross-correlation function is depicted in red. Peak lag [CI]: -3 days [-8,1].

Fig 11. Cross-correlation sequence estimate between SOREU regional incoming calls and daily infected. The blue lines represent the 90% confidence interval (CI) limits computed through a *z*-transformation. The maximum value of the cross-correlation function is depicted in red. Peak lag [CI]: -5 days [-11,1].

Fig 12. Cross-correlation sequence estimate between daily number of new tweets and daily infected. The blue lines represent the 90% confidence interval (CI) limits computed through a *z*-transformation. The maximum value of the cross-correlation function is depicted in red. Peak lag [CI]: -6 days [-8,-2].

In addition, a sensitivity analysis of these results with respect to the amplitude of the initial smoothing with a moving average filter was performed. The results are reported in Table 1.

Table 1. Sensitivity tests on the uncertainty in the location of the cross-correlation function peak. Different moving-average amplitudes have been used to test the robustness of the results of the confidence intervals (C.I.) computed through the *z*-transformation.

	NUE calls		**SOREU calls**		**tweets**	
n. days	**peak lag**	**C.I.**	**peak lag**	**C.I.**	**peak lag**	**C.I.**
4	-4	[-10,0]	-5	[-11,1]	-6	[-8,-3]
5	-4	[-9,0]	-5	[-11,1]	-6	[-8,-3]
6	-4	[-9,1]	-5	[-11,1]	-5	[-8,-2]

7	-3	[-8,1]	-5	[-11,1]	-6	[-8,-2]
8	-3	[-8,1]	-4	[-11,1]	-5	[-8,-2]
9	-2	[-8,1]	-4	[-10,0]	-5	[-8,-2]
10	-2	[-8,1]	-3	[-10,1]	-5	[-8,-2]

Changes in the results when the amplitude of the window is varying are relatively small (Table 1), consequently it can be assumed that our results are robust with respect to this parameter.

Similar and consistent results have been obtained with the Monte Carlo method described in the Supplementary material (Figs 13-15).

Fig 13. Cross-correlation analysis through a Monte Carlo simulation method for the regional NUE daily incoming calls vs. the daily infected time series. A random phase test (see Supplementary material) has been performed (1,000 simulations) to compute the time lag to "align" the signals and the corresponding confidence interval (C.I.): time lag = -4 days (C.I. -11,1).

Fig 14. Cross-correlation analysis through a Monte Carlo simulation method for the regional SOREU daily incoming calls vs. the daily infected time series. A random phase test (see Supplementary material) has been performed (1,000 simulations) to compute the time lag to "align" the signals and the corresponding confidence interval (C.I.): time lag = -6 days (C.I. -13,1).

Fig 15. Cross-correlation analysis through a Monte Carlo simulation method for the daily new tweets vs. the daily infected time series. A random phase test (see Supplementary material) has been performed (1,000 simulations) to compute the time lag to "align" the signals and the corresponding confidence interval (C.I.): time lag = -3 days (C.I. -11,4).

Discussion

Lombardy region has been the epicentre of the CoViD-19 epidemic in the Western Countries [28]. After the detection of the first case, on February the 20th, national and regional health authorities put in place several strategies to limit the spread of the infection and deal with the consequences of the increasing number of cases [29-31]. Indicators that can reveal and anticipate a rise of cases are of paramount importance to support the planning and interventions of the health service organization. Our study shows that the number of calls to emergency services could be a good indicator that can anticipate the need for hospitalization. However, some considerations about the analysis we conducted seem to be appropriate. With respect to wavelet analysis, wavelet decomposition in the time-scale plane has the advantage of giving a huge, precise and detailed amount of information both about time localization and frequency components. Nonetheless, this kind of analysis has also some limitations. First of all, time and frequency resolutions, according to uncertainty Heisenberg principle, are inversely proportional. Secondly, the edge effects affect the reliability of the results in the time-scale plane outside the cone of influence (see Supplementary material for an explanation), so that for the lowest frequencies the time interval with reliable results is very short. Lastly, one must consider also the band-pass filtering action of the CWT: indeed, the frequency spectrum bounds are a function of the number of samples of the signal (see Supplementary material for a detailed explanation), so the availability of data over time could represents a limitation for this kind of analysis. In our specific instance, it can be noticed that the inferior limit of the frequency domain in the case of Twitter data is much lower than the one of the daily regional emergency calls time series (112, 118, toll-free number), just because much more samples for the Twitter data are available (Figs 1-7). Consequently, in the cases of the calls to the emergency services, the ability of wavelet decomposition in revealing hidden signals and trends at the lowest frequencies is limited by the poor number of collected data. This limitation also affects

the WCS/MSWC computation and the possibility of detecting large time delays between the signals. This occurs because the lowest frequencies are cut off by the wavelet band-pass filter, so the duration of a cycle corresponding to the inferior limit of the frequency spectrum (i.e. the longest period of the band) is shorter than the time delay between the two signals. Thus, the large anticipation in the ascending phase evident from the time courses (see in the results the delays between the peak days) cannot be captured by WCS/MSWC. A future perspective would be to obtain and analyze a more complete set of data over the time scale: as already explained, with a greater number of samples, indeed, the minimum spectrum frequency gets lower, the corresponding duration of a cycle (i.e. the maximum period) increases and consequently wavelet analysis through CWT could be able to detect larger time delays between the signals. Moreover, even if aggregated regional data allow us to capture the "sum" of the effects of different local situations with probable a greater anticipation capability, if we consider the enormous geographical heterogeneity of CoViD-19 spread, regional data possess a limited usefulness for public health monitoring and preparedness with respect to a possible second wave of the epidemic. One more future direction would be to analyze these data at a more local level, such as for each SOREU and each province (NUTS-3 level), or even municipalities, of Lombardy region. To supply the lack of generality of wavelet analysis, we performed also a time lag analysis through the cross-correlation function. Once again, the limited number of data affected our results: indeed, while the peaks of the emergency calls anticipated that of daily new cases by about two weeks (Figs 2-3), the delay of the cross-correlation function peak is much lower both for NUE calls (-3 days and -4 days, respectively for the original data and the modelled time series, see Figs 10 and 13) and for SOREU calls (-5 days and -6 days, respectively for the original data and the surrogate time series, see Figs 11 and 14). This can be explained just looking at the time courses (Figs 2-3): the availability of emergency calls data is such that we can observe a great part of the descending phase, whereas for the infected curve an initial decrease is visible only during the last days of the corresponding observation period. With this perspective, consequently, it would be useful to analyze a more complete dataset. Several other

indicators are currently under investigation and many of them will provide useful information, but we should not only rely on indicators focused on detecting an increase in new cases, because the main impact on the health system is more related to the characteristics of the infected population rather than to the number of infected people.

The severe countermeasures put in place, such as the national lockdown, had a deep impact on the population from several points of view, not only on the health system. It is therefore important to take into account the social reaction to the crisis and analyzing it is part of the public health response. Our analysis shows that Twitter trends correlate more with social factors rather than with the number of cases (Figs 8-9). This finding suggests that a thorough analysis of social media would improve our understanding about what the most common worries, fears and feelings of the population are, in order to address them through a public health strategy that should include a proper use of social media to inform the population. Among all the Twitter data, only the daily number of new tweets reveals some anticipation capability with respect to the epidemic curve: wavelet analysis, indeed, detects a trend at the lowest frequencies, and a phase-lagged anomaly in a frequency range centred around 0.25 cycles/day that occurs just between the two peaks (Fig 4). This finding is confirmed and consistent both with the 7-days distance of the two curve peaks and with the cross-correlation analysis (maximum at -6 days and -3 days lags, respectively for the original data and the modelled time series, with small confidence intervals, see Figs 12 and 15).

References

1. Cereda, D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, et al. The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv:2003.09320v1 [Preprint]. 2020 [submitted 2020 March 20; cited 2020 October 5]. Available from: https://arxiv.org/abs/2003.09320v1

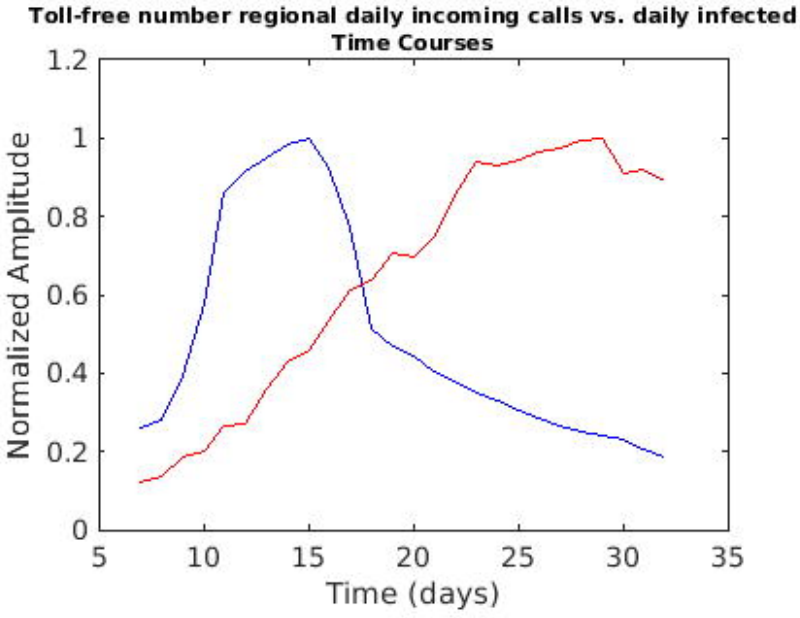
2. Rivieccio BA, Luconi E, Boracchi P, Pariani E, Romanò L, Salini S, et al. Heterogeneity of COVID-19 outbreak in Italy. Acta Biomed. 2020; 91:31-4.
3. Italian Department for Civil Defense [Internet]. CoViD-19 Italia. Monitoraggio situazione.[cited 2020 October 5]. Available from: https://github.com/pcm-dpc/COVID-19. Italian.
4. WHO. Novel Coronavirus (2019-nCoV) situation report - 13. 2020 Feb 2. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf
5. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015;43(6):563-71.
6. Castaldi S, Maffeo M, Rivieccio BA, Zignani M, Manzi G, Nicolussi F, et al. Monitoring emergency calls and social networks for COVID-19 surveillance. To learn for the future: the outbreak experience of the Lombardia region in Italy. Acta Biomed. 2020;91:29-33.
7. vita.it [Internet]. Coronavirus, numeri di emergenza presi d'assalto. [posted 2020 February 24; cited 2020 October 5]. Available from: http://www.vita.it/it/article/2020/02/24/coronavirus-numeri-di-emergenza-presi-dassalto/154125/. Italian.
8. EENA. EENA recommendations for Emergency Services Organisations during the COVID-19 outbreak. Available from: https://eena.org/knowledge-hub/documents/eena-recommendations-for-emergency-services-organisations-during-the-covid-19-outbreak/
9. EENA. Appendix. Data and strategies per country on emergency calls & public warning during COVID-19 outbreak. Available from: https://eena.org/knowledge-hub/documents/data-and-strategies-per-country-on-emergency-calls-public-warning-during-covid-19
10. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Health Surveill [Internet]. 2020 Apr-Jun [cited 2020 October 5];6(2):e19723. Available from: https://publichealth.jmir.org/2020/2/e19273/

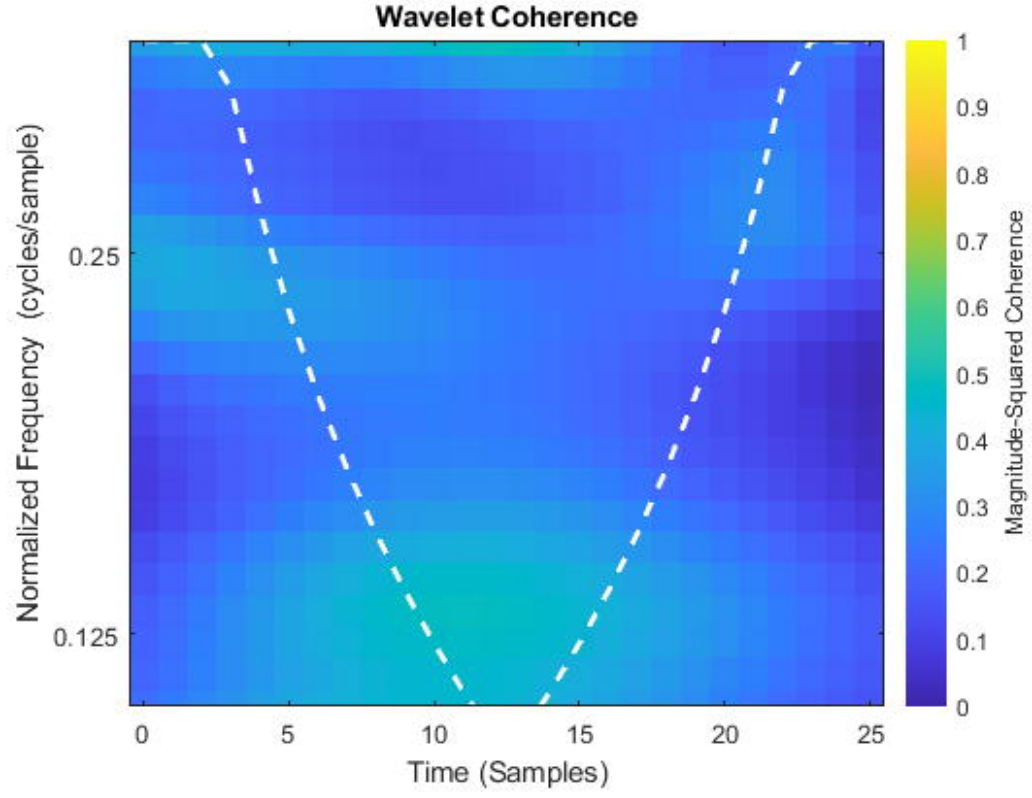
11. Sooknanan J, Comissiong DMG. Trending on social media: integrating social media into infectious disease dynamics. Bull Math Biol. 2020;82:86.
12. twitter.com [Internet]. The COVID tracking project. [cited 2020 October 5]. Available from: https://twitter.com/covid19tracking
13. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. J Biomed Inform. 2020;108:103500.
14. Castaldi S, Romanò L, Pariani E, Garbelli C, Biganzoli E. COVID-19: the end of lockdown what next? Acta Biomed. 2020;91:236-8.
15. ilsole24ore.com [Internet]. Chiamate al 118, in Lombardia oltre il 30% per motivi respiratori e infettivi. Picco il 16 marzo. [posted 2020 April 7; cited 2020 October 5]. Available from: https://www.ilsole24ore.com/art/chiamate-118-lombardy-oltre-30percento-motivi-respiratori-e-infettivi-picco-16-marzo-ADEYHnI?refresh_ce=1. Italian.
16. ilsole24ore.com [Internet]. In Lombardia calano le telefonate al 112, il picco il 12 marzo. Resta elevato il rapporto con i ricoveri. [posted 2020 April 4; cited 2020 October 5]. Available from: https://www.ilsole24ore.com/art/in-lombardy-calano-telefonate-112-picco-13-marzo-resta-elevato-rapporto-i-ricoveri-ADp1r3H. Italian.
17. twitter.com [Internet]. [cited 2020 October 5]. Available from: https://twitter.com/search?q=(118%20OR%20112)%20lang%3Ait%20until%3A2020-06-29%20since%3A2020-02-18&src=typed_query
18. Bali R, Sarkar D, Lantz B, Lesmeister C. R: unleash machine learning techniques. Birmingham (UK): Packt Publishing Ltd; 2016.
19. Daubechies I. The wavelet transform time-frequency localization and signal analysis. IEEE Trans Inform Theory. 1990;36:961-1004.
20. Stoica P, Moses R. Spectral analysis of signals. Upper Saddle River (NJ): Prentice Hall; 2005.
21. Tian F, Tarumi T, Liu H, Zhang R, Chalak L. Wavelet coherence analysis of dynamic cerebral autoregulation in neonatal hypoxic-ischemic encephalopathy. Neuroim Clin. 2016;11:124-32.

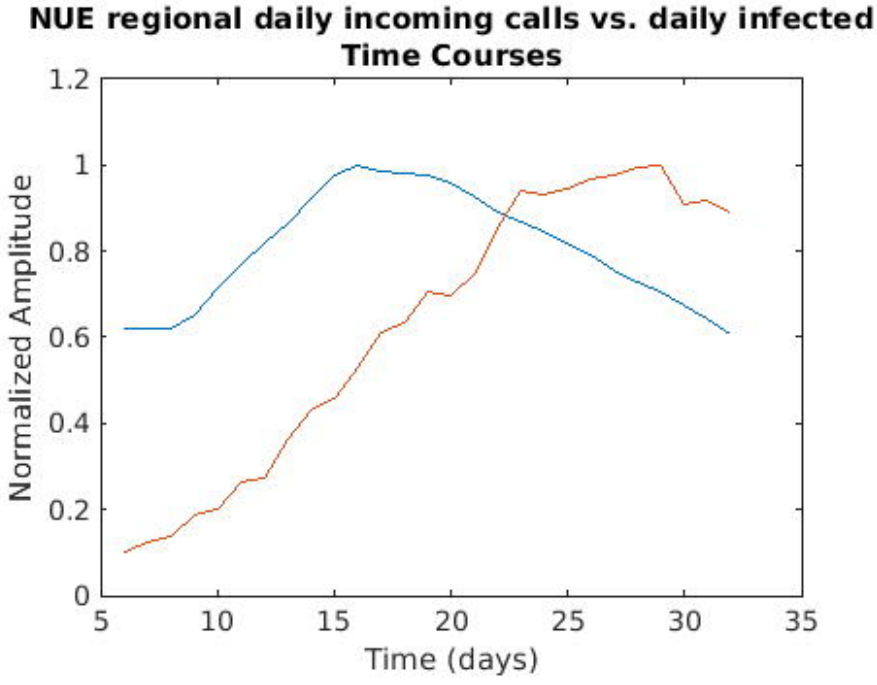
22. Torrence C, Compo GP. A practical guide to wavelet analysis. Bull Am Metereol Soc. 1998;79:61-78.
23. Ogden T. Essential wavelets for statistical applications and data analysis. Basel (CH): Birkhäuser; 1997.
24. Percival DB, Walden AT. Wavelet methods for time series analysis. Cambridge (UK): Cambridge University Press; 2000.
25. Meyers SD, Kelly BG, O'Brien JJ. An introduction to wavelet analysis in oceanography and meteorology: with application to the dispersion of Yanai waves. Monthly Weather Rev. 1993;121:2858-66.
26. Fisher RA. On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron. 1921;1:3-32.
27. Ebisuzaki W. A method to estimate the statistical significance of a correlation when the data are serially correlated. J Climate. 1997;10(9):2147-53.
28. euronews.com [Internet]. Coronavirus: more than a third of people in Italy's COVID-19 epicentre estimated to have had disease. [posted 2020 April 28; cited 2020 October 5]. Available from: https://www.euronews.com/2020/04/28/coronavirus-more-than-a-third-of-people-in-italy-s-covid-19-epicentre-estimated-to-have-ha
29. Disposizioni attuative del decreto-legge 23 febbraio 2020, n. 6, recante misure urgenti in materia di contenimento e gestione dell'emergenza epidemiologica da COVID-19, Italian Prime Minister Decree, Gazzetta Ufficiale Serie Generale n. 45 (Feb 23, 2020). Italian.
30. Misure straordinarie ed urgenti per contrastare l'emergenza epidemiologica da COVID-19 e contenere gli effetti negativi sullo svolgimento dell'attività giudiziaria. Italian Law Decree n. 11, Gazzetta Ufficiale Serie Generale n. 60 (Mar 8, 2020). Italian.
31. Costituzione dell'Unità di Crisi dell'emergenza sanitaria in ordine all'emergenza epidemiologica da COVID-19 e relativa Task Force. Lombardy Region Secretariat Decree n.

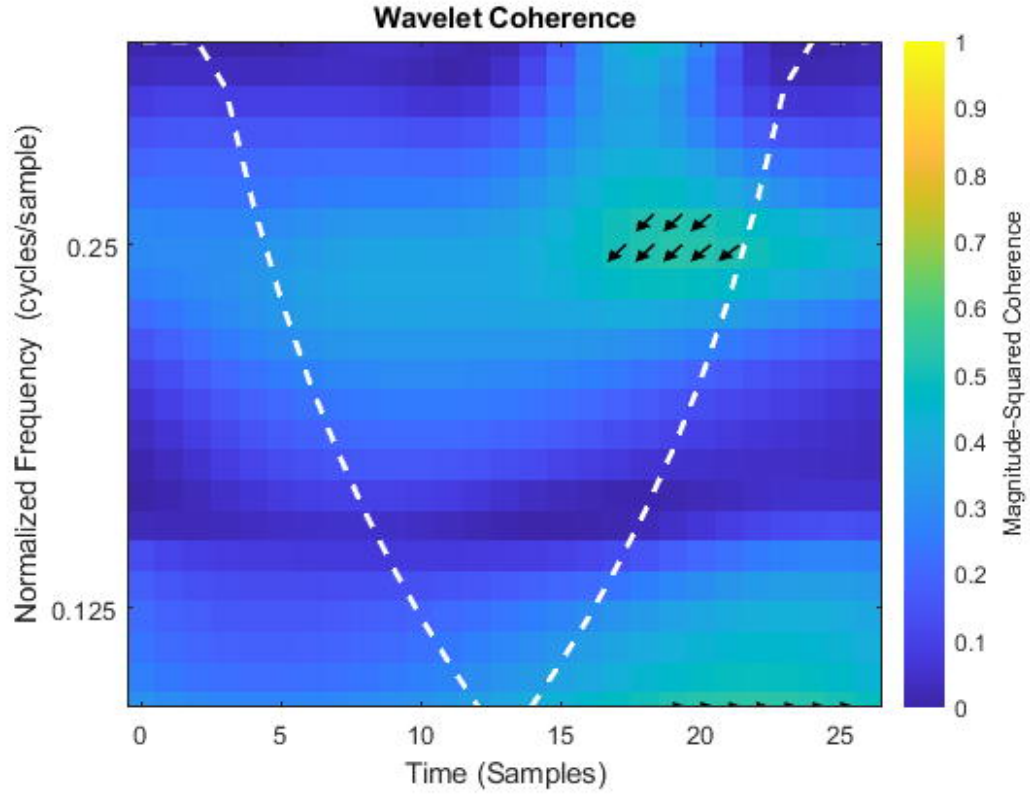
3287 (Mar 12, 2020). Available from: https://www.openpolis.it/wp-content/uploads/2020/05/DECRETO-3287-DEL-12NMARZO-2020.pdf. Italian.

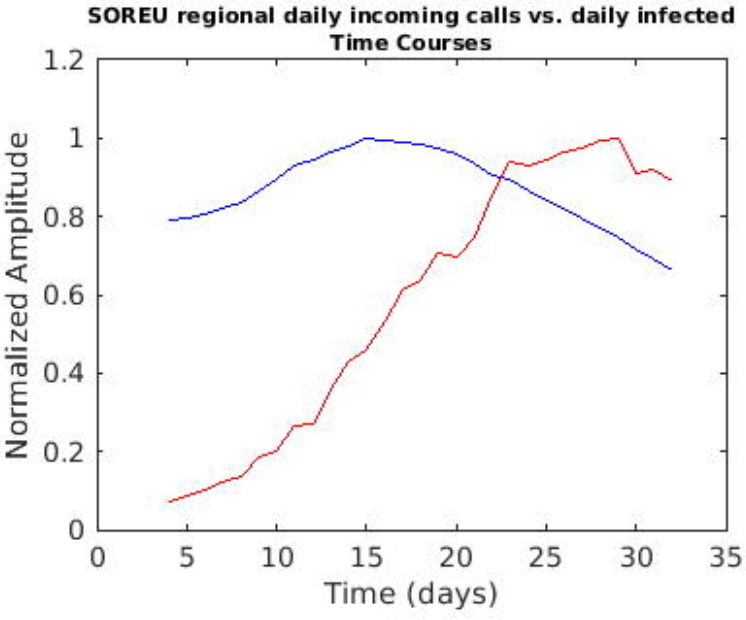
32. Dziedziech K, Nowak A, Hasse A, Uhl T, Staszewski WJ. Wavelet-based analysis of time-variant adaptive structures. Philos Trans A Math Phys Eng Sci. 2018;13:376(2126):20170245.
33. Grinsted A, Moore JC, Jevrejeva S. Application of the cross-wavelet transform and wavelet coherence to geophysical time series. Nonlin Proc Geophys. 2004;11:561-6.
34. Maraun, Kurths DJ, Holschneider M. Nonstationary Gaussian processes in wavelet domain: synthesis, estimation and significance testing. Phys Rev E. 2007;75(1 Pt 2):016707.
35. Torrence C, Webster P. Interdecadal changes in the ESNO-Monsoon System. J Clim. 1999;12:2679-90.

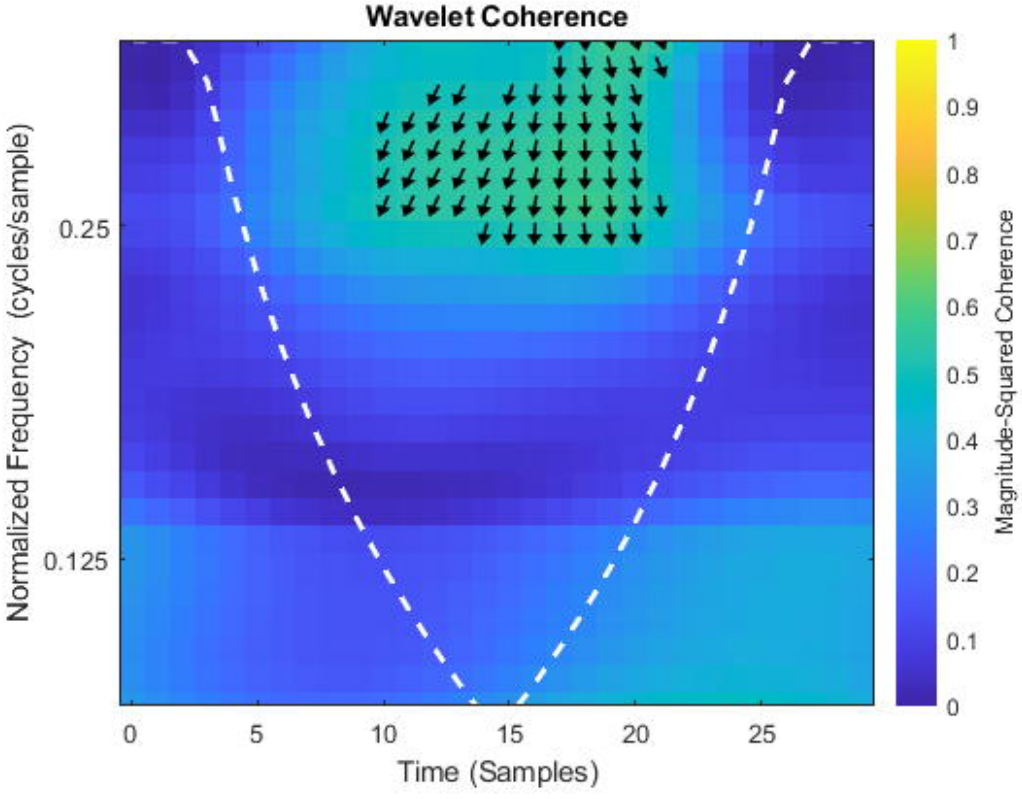

Toll-free number regional daily incoming calls vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
5
10
15
20
25
30
35

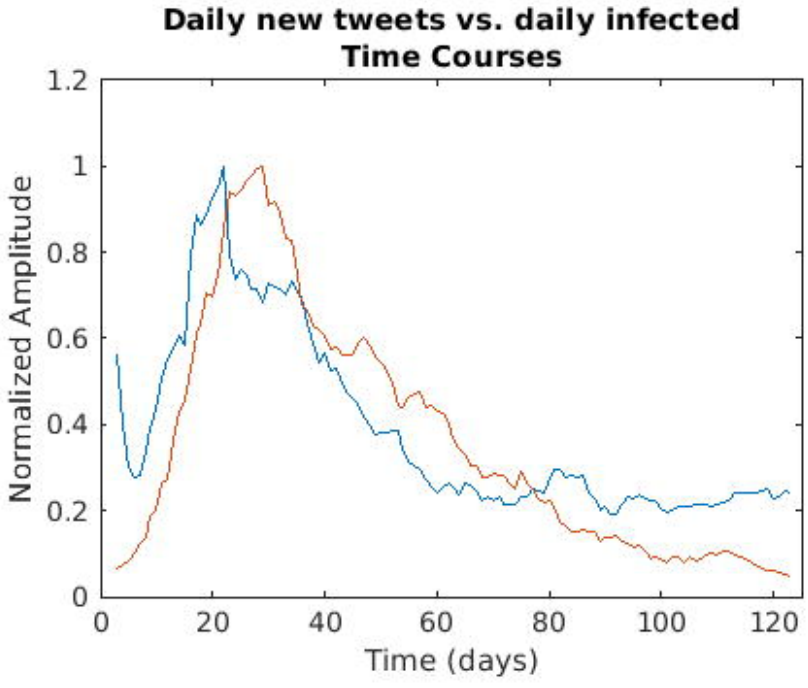

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
Time (Samples)
0
5
10
15
20
25
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

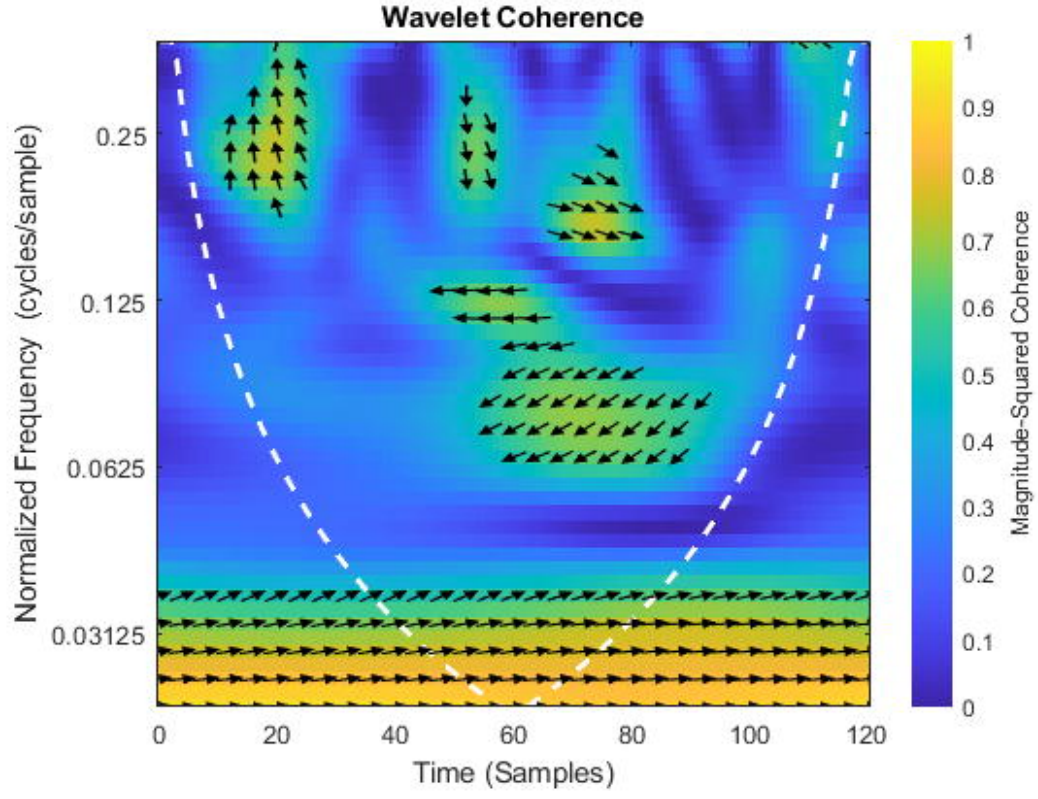

NUE regional daily incoming calls vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
5
10
15
20
25
30
35

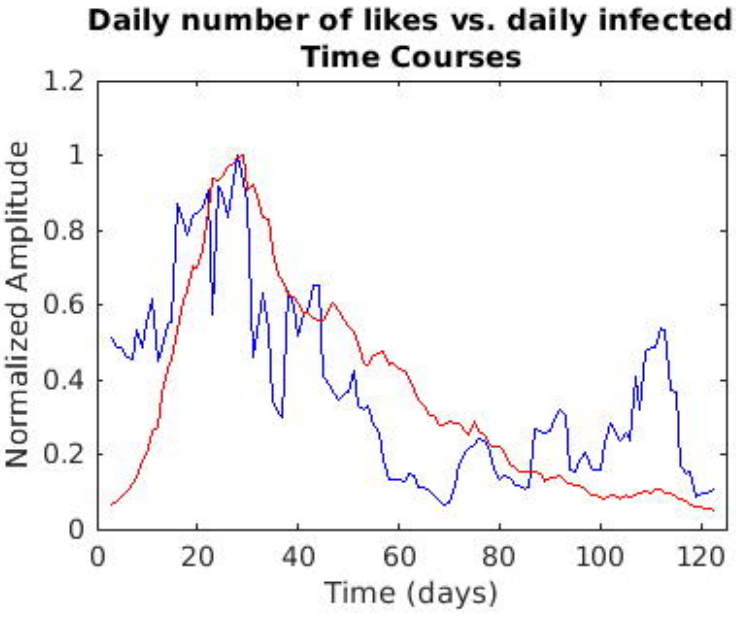

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
Time (Samples)
0
5
10
15
20
25
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

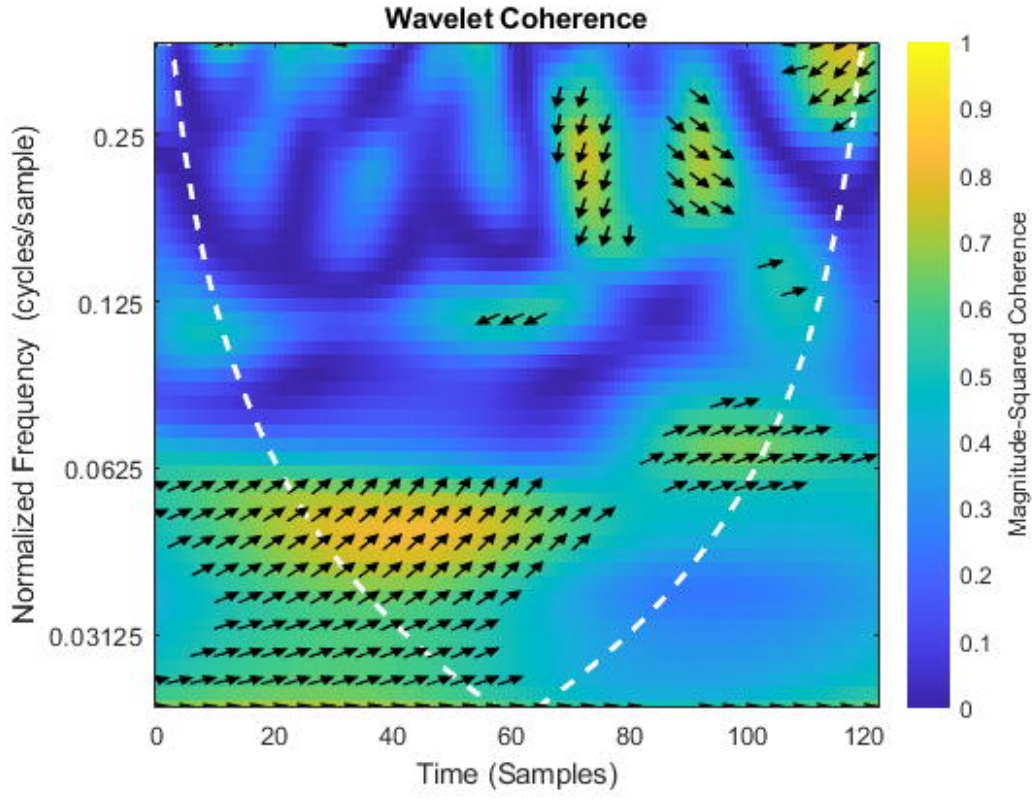

SOREU regional daily incoming calls vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
0
5
10
15
20
25
30
35

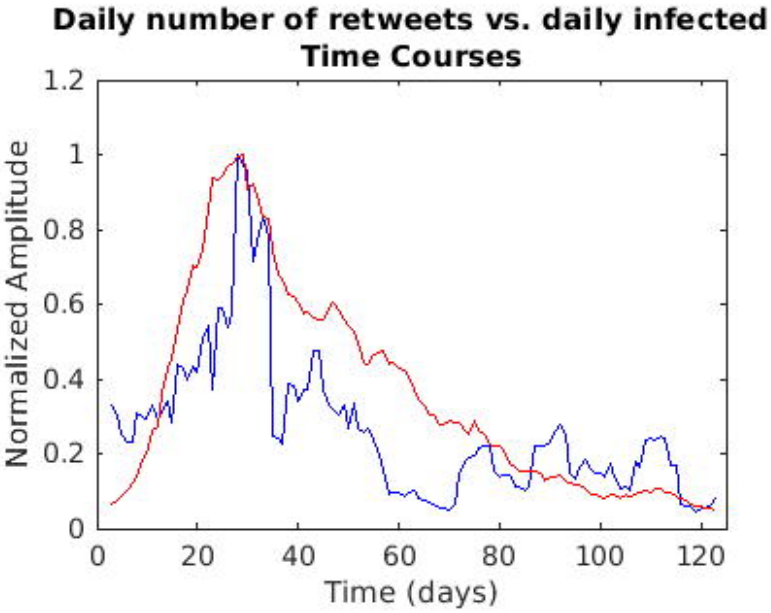

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
Time (Samples)
0
5
10
15
20
25
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

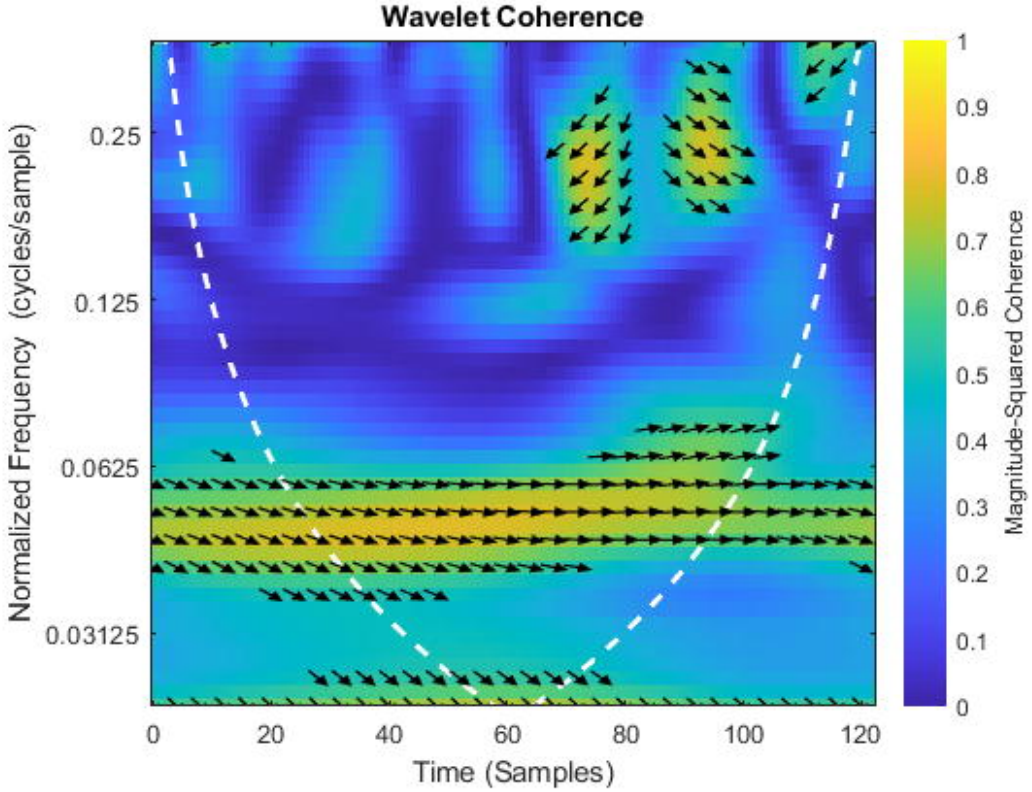

Daily new tweets vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
0
20
40
60
80
100
120

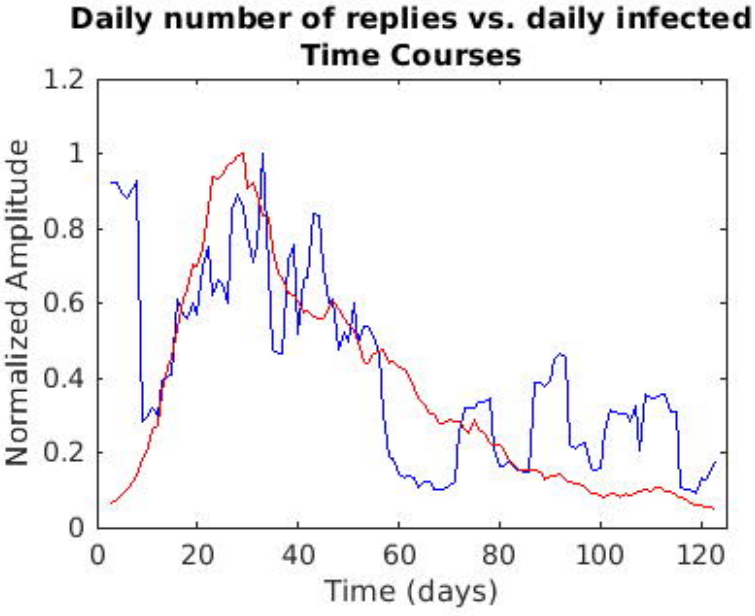

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
0.0625
0.03125
Time (Samples)
0
20
40
60
80
100
120
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

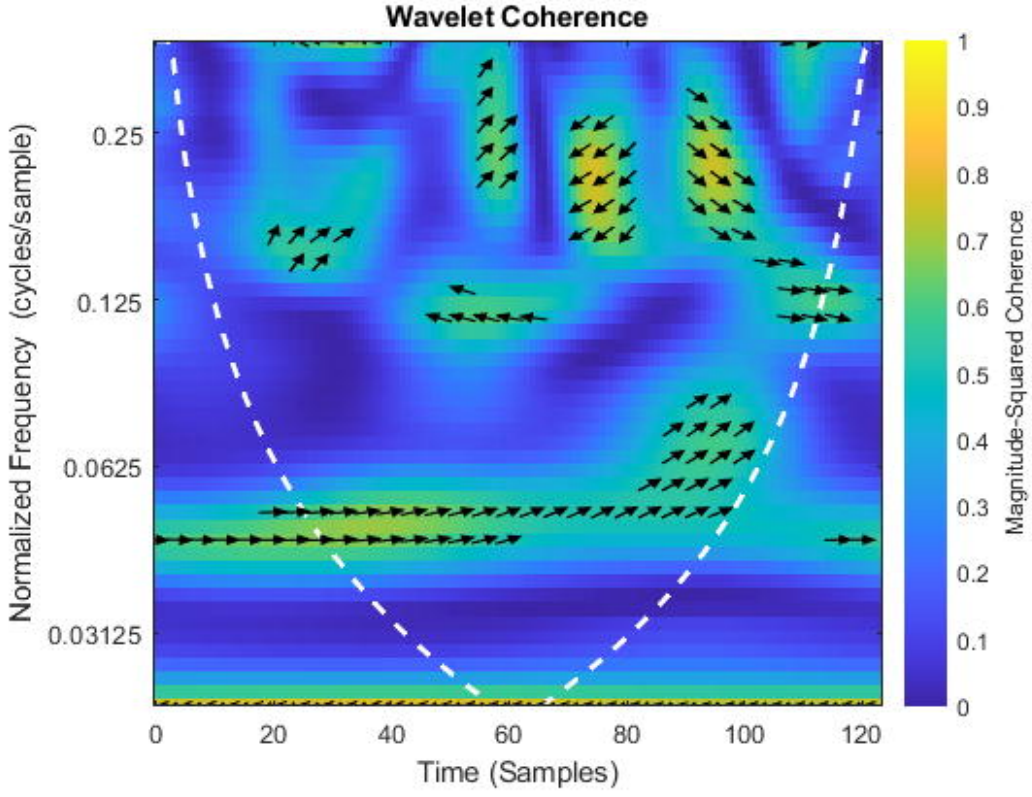

Daily number of likes vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
0
20
40
60
80
100
120

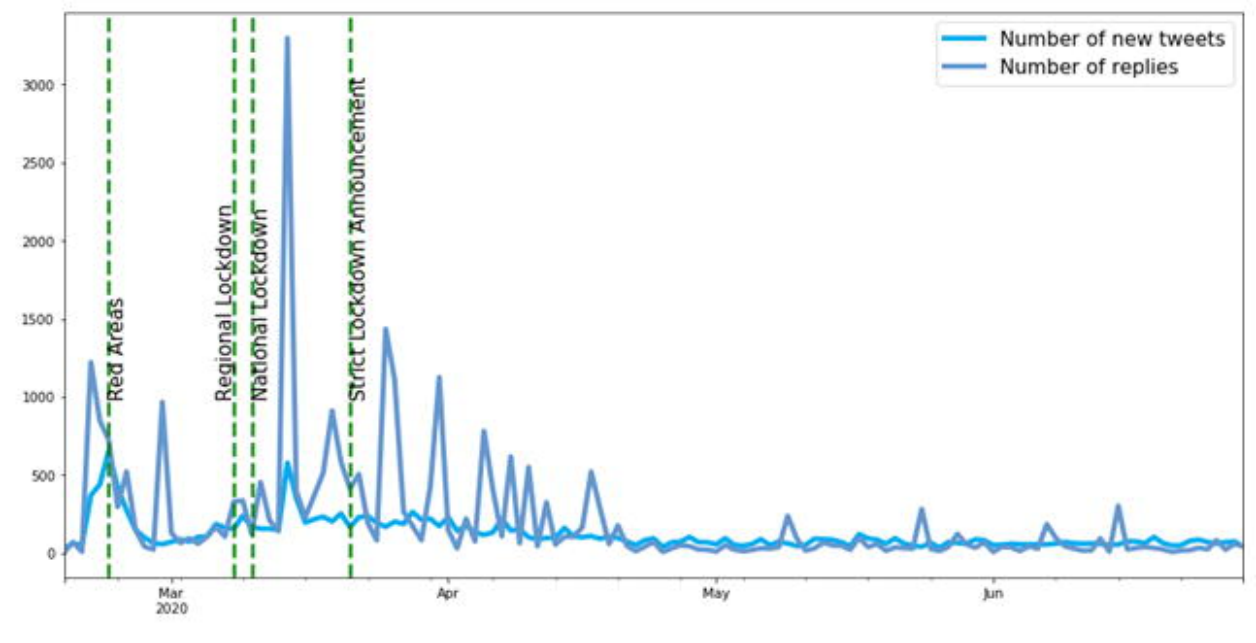

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
0.0625
0.03125
Time (Samples)
0
20
40
60
80
100
120
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

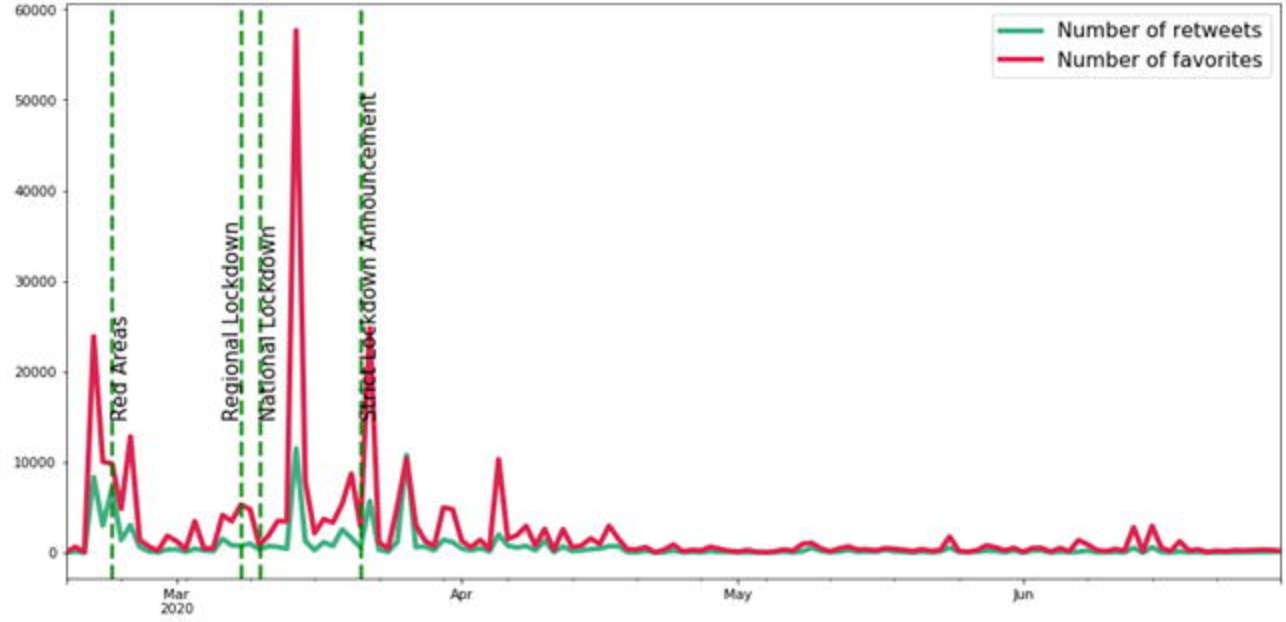

Daily number of retweets vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
0
20
40
60
80
100
120

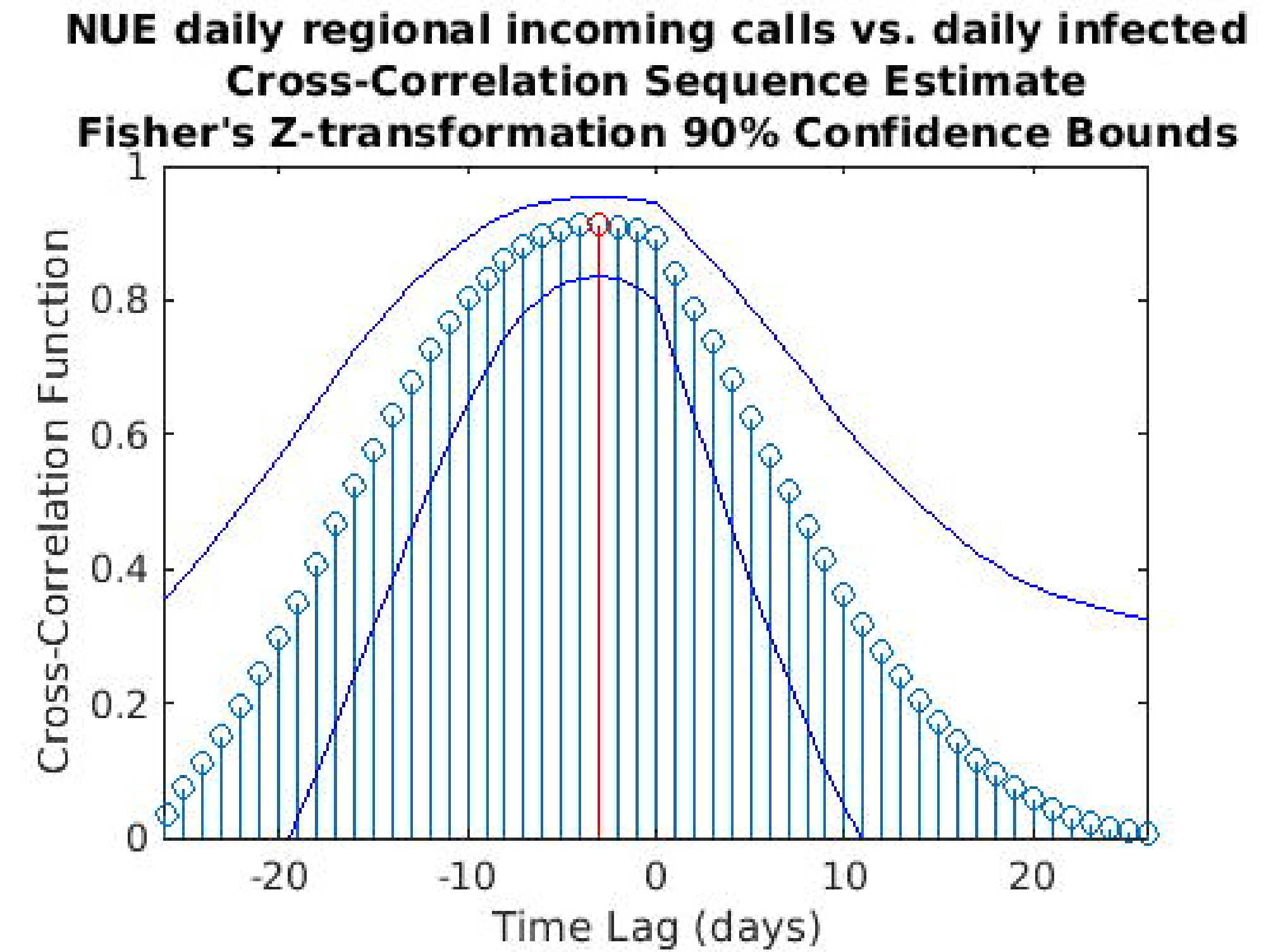

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
0.0625
0.03125
Time (Samples)
0
20
40
60
80
100
120
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

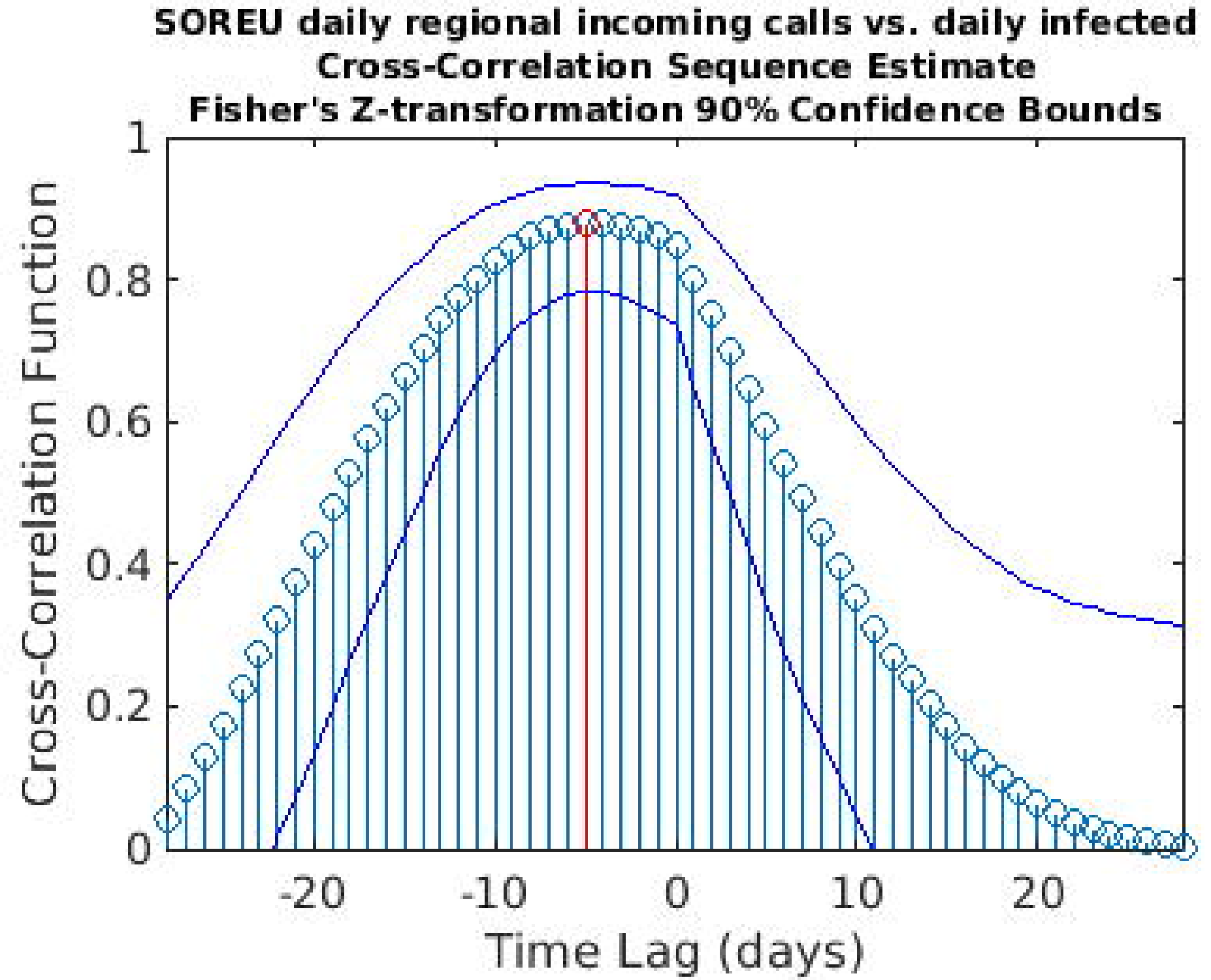

Daily number of replies vs. daily infected
Time Courses
Normalized Amplitude
Time (days)
0
0.2
0.4
0.6
0.8
1
1.2
0
20
40
60
80
100
120

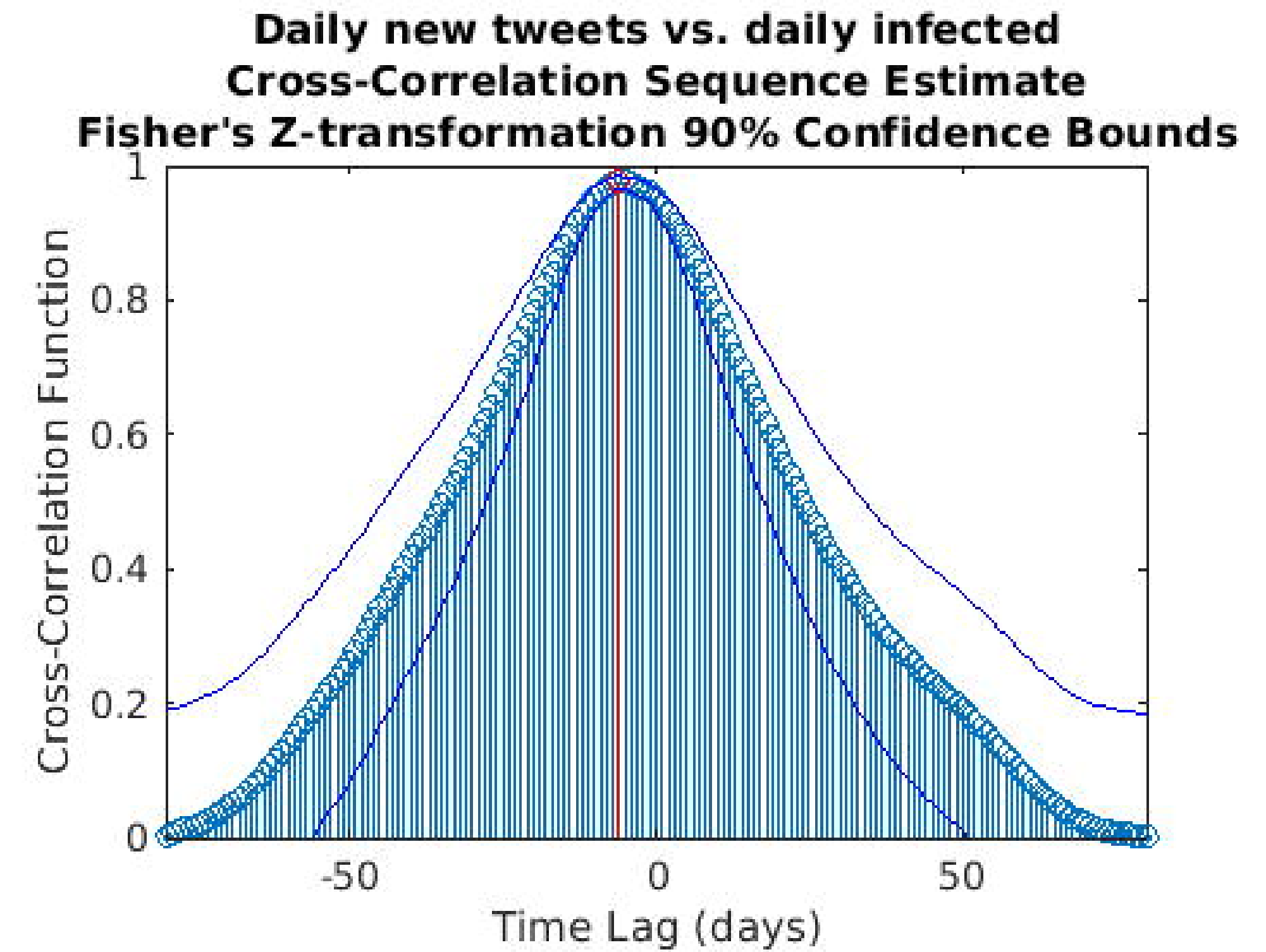

Wavelet Coherence
Normalized Frequency (cycles/sample)
0.25
0.125
0.0625
0.03125
Time (Samples)
0
20
40
60
80
100
120
Magnitude-Squared Coherence
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

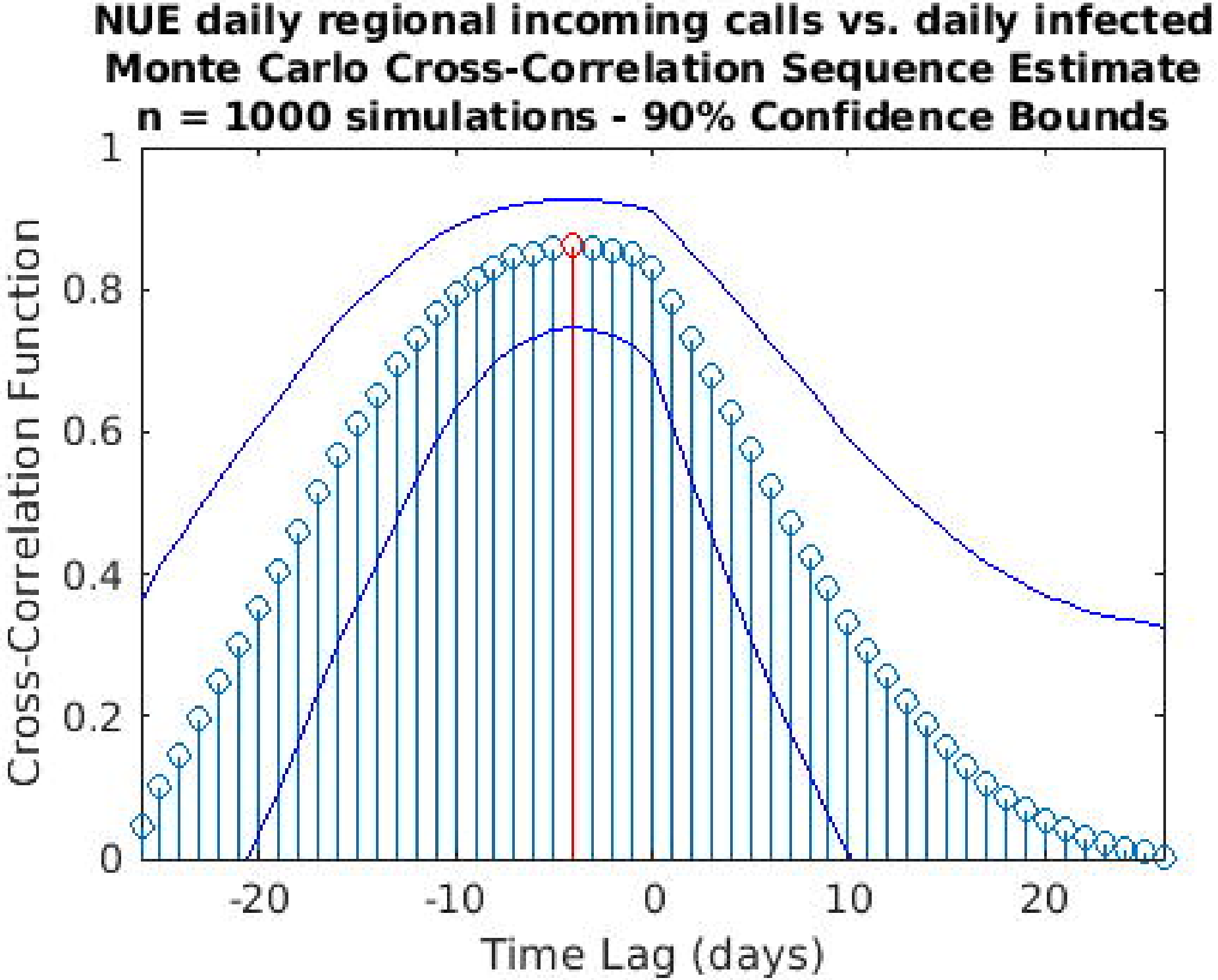

Number of new tweets
Number of replies
Red Areas
Regional Lockdown
National Lockdown
Strict Lockdown Announcement
3000
2500
2000
1500
1000
500
0
Mar
2020
Apr
May
Jun

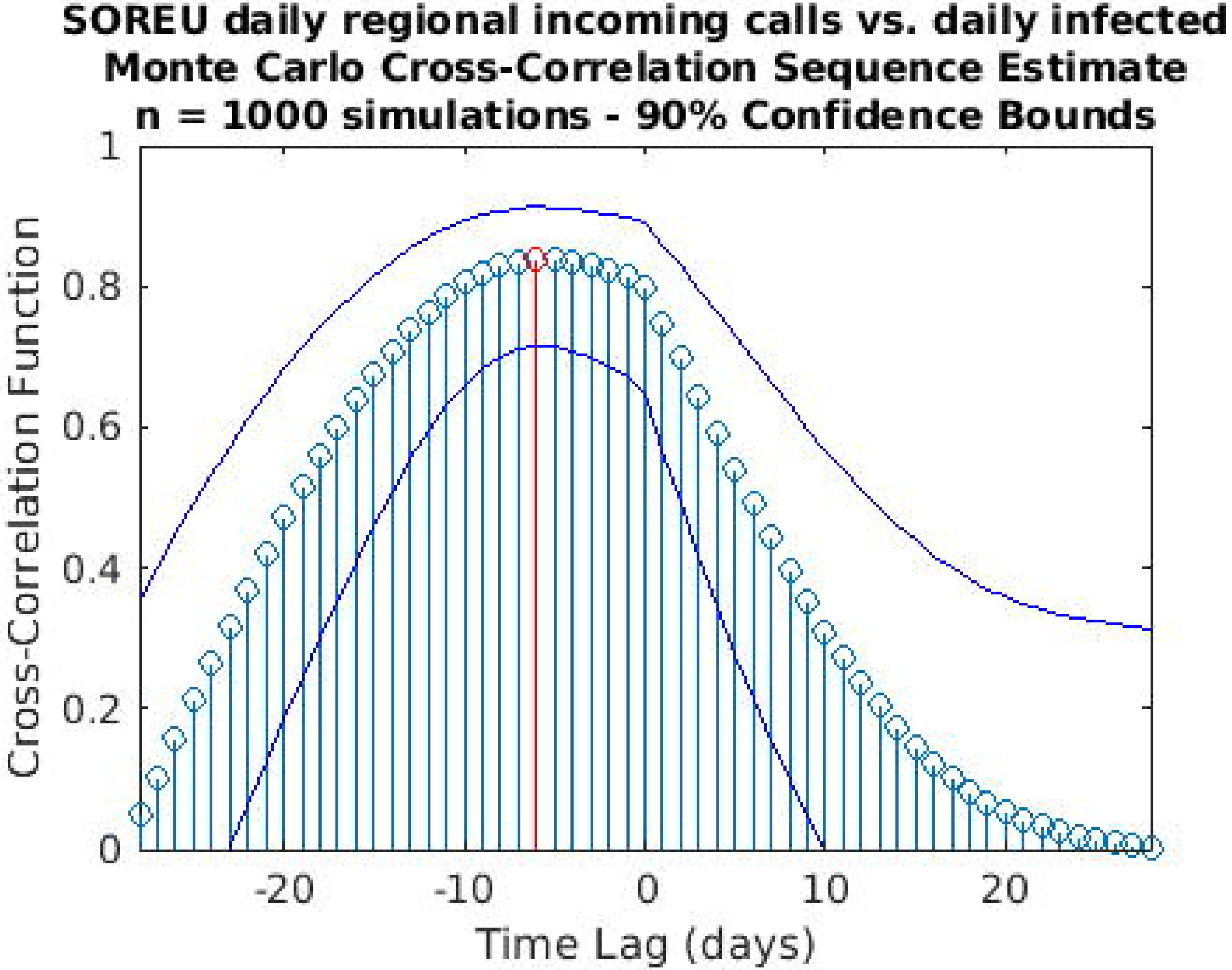

Number of retweets
Number of favorites
Red Areas
Regional Lockdown
National Lockdown
Strict Lockdown Announcement
60000
50000
40000
30000
20000
10000
0
Mar
2020
Apr
May
Jun

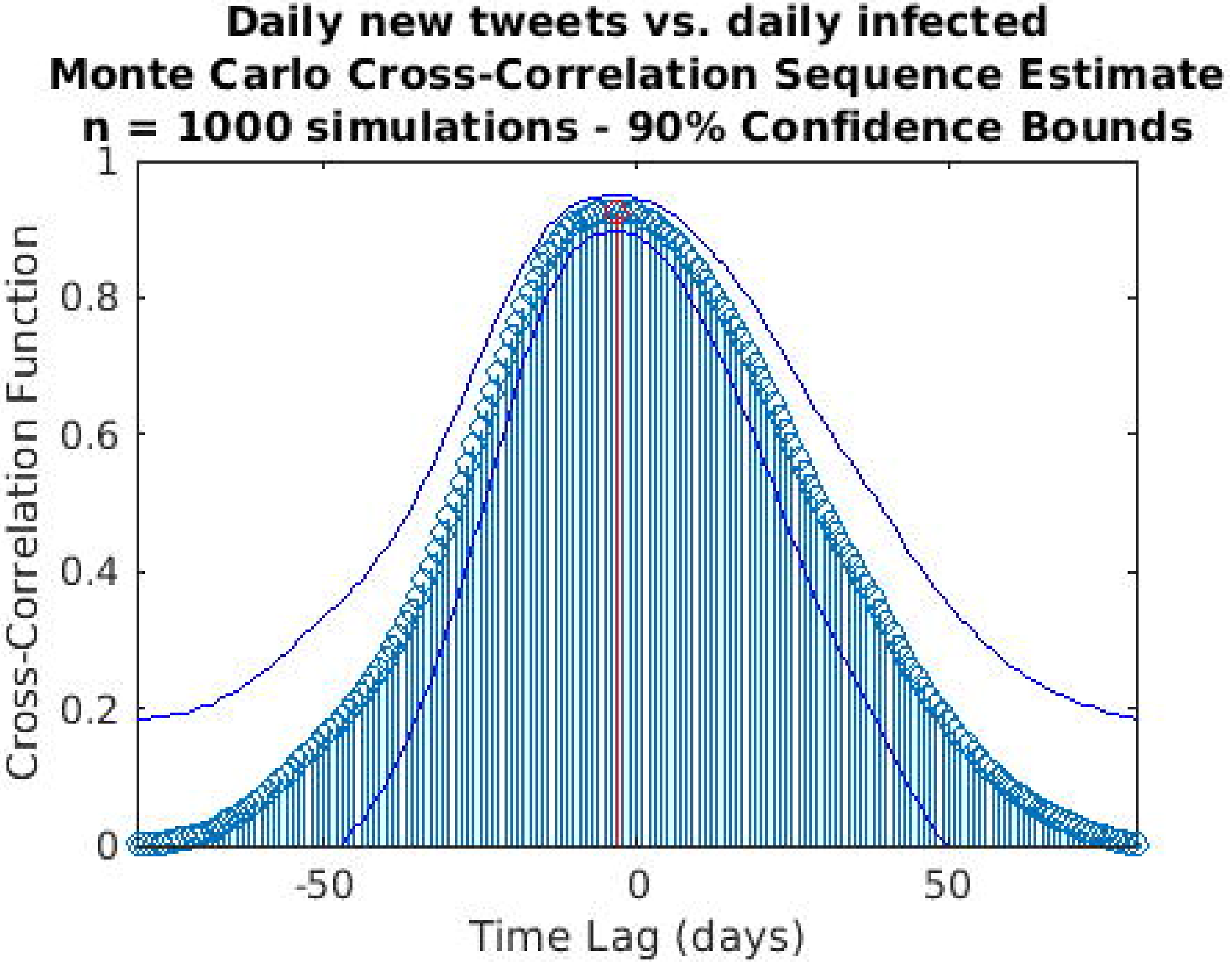

NUE daily regional incoming calls vs. daily infected
Cross-Correlation Sequence Estimate
Fisher's Z-transformation 90% Confidence Bounds
Cross-Correlation Function
Time Lag (days)
1
0.8
0.6
0.4
0.2
0
-20
-10
0
10
20


SOREU daily regional incoming calls vs. daily infected
Cross-Correlation Sequence Estimate
Fisher's Z-transformation 90% Confidence Bounds
Cross-Correlation Function
Time Lag (days)
1
0.8
0.6
0.4
0.2
0
-20
-10
0
10
20


Daily new tweets vs. daily infected
Cross-Correlation Sequence Estimate
Fisher's Z-transformation 90% Confidence Bounds
Cross-Correlation Function
1
0.8
0.6
0.4
0.2
0
-50
0
50
Time Lag (days)


NUE daily regional incoming calls vs. daily infected
Monte Carlo Cross-Correlation Sequence Estimate
n = 1000 simulations - 90% Confidence Bounds
Cross-Correlation Function
Time Lag (days)
1
0.8
0.6
0.4
0.2
0
-20
-10
0
10
20


SOREU daily regional incoming calls vs. daily infected
Monte Carlo Cross-Correlation Sequence Estimate
n = 1000 simulations - 90% Confidence Bounds
Cross-Correlation Function
Time Lag (days)
1
0.8
0.6
0.4
0.2
0
-20
-10
0
10
20


Daily new tweets vs. daily infected
Monte Carlo Cross-Correlation Sequence Estimate
n = 1000 simulations - 90% Confidence Bounds
Cross-Correlation Function
Time Lag (days)
1
0.8
0.6
0.4
0.2
0
-50
0
50