

A New Data Integration Framework for Covid-19 Social Media Information

Lauren Ansell and Luciana Dalla Valle*

University of Plymouth

October 11, 2021

Abstract

The Covid-19 pandemic presents a serious threat to people's health, resulting in over 250 million confirmed cases and over 5 million deaths globally. In order to reduce the burden on national health care systems and to mitigate the effects of the outbreak, accurate modelling and forecasting methods for short- and long-term health demand are needed to inform government interventions aiming at curbing the pandemic. Current research on Covid-19 is typically based on a single source of information, specifically on structured historical pandemic data. Other studies are exclusively focused on unstructured online retrieved insights, such as data available from social media. However, the combined use of structured and unstructured information is still uncharted. This paper aims at filling this gap, by leveraging historical as well as social media information with a novel data integration methodology. The proposed approach is based on vine copulas, which allow us to improve predictions by exploiting the dependencies between different sources of information. We apply the methodology to combine structured datasets retrieved from official sources and to a big unstructured dataset of information collected from social media. The results show that the proposed approach, compared to traditional approaches, yields more accurate estimations and predictions of the evolution of the Covid-19 pandemic.

Keywords: Covid-19; Dependence Modelling; Google Trends; Social Media Sentiment Analysis; Time Series Modelling; Twitter.

*Corresponding author. E-mail: luciana.dallavalle@plymouth.ac.uk. Address: Drake Circus, PL48AA, Plymouth (UK)

1 Introduction

The outbreak of the Covid-19 disease has infected and killed millions of people globally, resulting in a pandemic with enormous global impact. This disease affects the respiratory system, and the viral agent that causes it spreads through droplets of saliva, as well as through coughing and sneezing. As an extremely transmissible viral infection, Covid-19 is causing significant damage to countries' economies because of its direct impact on the health of citizens and the containment measures taken to curtail the virus (Pinheiro et al., 2021). In the UK, Covid-19 had serious implications for people's health and the healthcare services, with more than 8 million confirmed cases and 150,000 deaths, as government figures show. There is thus a widespread interest in accurately estimating and assessing the evolution of the pandemic over time.

Current studies on Covid-19 are typically based on a single source of information. Most of them implement quantitative analyses focusing on historical data to produce forecasts of the pandemic. For example, Li et al. (2020) used clinical data of Covid-19 patients to statistically analyse with meta-analysis the clinical symptoms and laboratory results aiming at explaining the discharge and fatality rate. Rahimi et al. (2021) considered data on confirmed and recovered cases and deaths, the growth rate and the trend of the disease in Australia, Italy and the UK. The authors predicted epidemiology in the three countries with mathematical approaches based on susceptible, infected, and recovered (SIR) cases and susceptible, exposed, infected, quarantined, and recovered (SEIQR) cases, comparing them with the Prophet machine learning algorithm and the logistic regression. Machine learning methods were also adopted, for example, by DeCaprio et al. (2020), who implemented logistic regression and gradient boosted trees on health risk assessment questionnaires and medical claims diagnosis data to predict complications due to Covid-19. Ahmed et al. (2021) performed descriptive, diagnostic, predictive, and prescriptive analysis of the pandemic applying neural networks and other machine learning algorithms, focusing on different pandemic symptoms. Pinheiro et al. (2021) employed network analytics and machine learning models by using a combination of anonymized health and telecommunications data to understand the correlation between population movements and virus spread and to predict possible new outbreaks.

However, some authors in the literature criticize approaches which are exclusively based on official quantitative information. Wynants et al. (2020) and Jewell et al. (2020) noted that existing Covid-19 studies based on historical data and prediction models for the pandemic are “poorly reported, at high risk of bias and underperforming”.

Another strand of research focuses on different types of data, analysing textual information with Natural Language Processing (NLP) methods. For example, Liu et al. (2021) adopted the Latent Dirichlet Allocation (LDA) model, to allocate research articles into different research topics pertinent to Covid-19 according to their abstracts. A LDA model was also employed by Wang et al. (2021), who applied it to the question and answer data about Covid-19 in Chinese online health communities. Text data in the form of social media insights are also used by other contributors in the literature to evaluate and predict the progression of the Covid-19 pandemic. For example, Li et al. (2020) employed a lag correlation analysis with data collected from Google Trends, Baidu Search Index and Sina Weibo Index. Sina Weibo messages were also analysed by Liu et al. (2020), Peng et al. (2020) and Zhu et al. (2020). The former authors carried out a statistical analysis based on the Fisher exact test, calculated the rates of death with the Kaplan-Meier method and established risk factors for mortality using the multivariate Cox regression. Peng et al. (2020) implemented a kernel density analysis and an ordinary least square regression to identify the spatiotemporal distribution of Covid-19 cases in the main urban area of Wuhan, China. Zhu et al. (2020) calculated descriptive statistics and applied a time series analysis to the data. Baidu Search Index information was analysed by Qin et al. (2020) using different statistical methods, including the subset selection, forward selection, lasso regression, ridge regression and elastic net. Google trends searches, Wikipedia page views and Twitter messages were gathered by O’Leary and Storey (2020), who implemented a regression analysis to show that online-retrieved information provided a leading indication of the number of people in the USA who became infected and die from the coronavirus.

However, contributions in the literature which are exclusively based on either official or online information as stand-alone sources are not taking into account the drawbacks affecting the data and the potential synergies between different data sources. On the one hand, due to limited capacity of testing, official data on confirmed cases are unlikely to reflect the true Covid-19 numbers. On the other hand, social media data are generated by users on a voluntary basis and may not capture information about the entire population. Therefore, predictive models built on a single source of information might generate biased results.

The goal of this paper is to develop a state-of-the-art data integration framework, leveraging the dependencies between historical and online data to provide more accurate evaluations and predictions of the Covid-19 dynamics. Our approach is based on vine copulas, which are very flexible mathematical tools, able to correctly capture the dependence structures between different variables (Czado, 2019). Integration of different data sources using copulas and Bayesian networks

was first proposed by Dalla Valle (2014) and, later, by Dalla Valle and Kenett (2015, 2018). However, the approach adopted by the authors was based on data calibration (Dalla Valle, 2017c). In this paper, our aim is to propose a comprehensive novel data integration framework, able to improve data modelling and forecasting, along the lines of the paper by Ansell and Dalla Valle (2021), who successfully applied a similar approach to small size natural hazard datasets.

So far, the application of copulas and vines to pandemic data has been limited to study the implications of Covid-19, especially in the financial field, rather than to directly calculate forecasts of pandemic trends. For example, Maneejuk et al. (2021) implemented a Markov-switching dynamic copula with Student-t distribution to explore Covid-19 shock effects on energy markets. Sifat et al. (2021) used a quantile regression model estimated via vine copula to show that speculation in energy and precious metal futures are more prevalent in crisis periods such as the Covid-19 pandemic.

This paper will be the first to propose statistical methodology based on vine copulas, able to exploit and integrate official and social media data, to accurately model the spread of Covid-19. The methodology will be applied to structured historical pandemic data and to a large unstructured online-retrieved dataset, relevant to the UK geographical area. The results show that our approach performs better than traditional approaches, which do not take into account associations between official and on-line information, to estimate and predict the Covid-19 dynamics.

The remainder of the paper is organised as follows. Section 2 describes the different data sources used in the analysis; Section 3 illustrates the vine copula methodology; Section 4 reports the results of the analysis; finally, concluding remarks are presented in Section 5.

2 Dataset

The structured and unstructured data used in this paper were collected daily between the 21st April 2020 and the 9th May 2021. As structured data, we considered the number of new admitted patients (**Admissions**), the number of hospital cases (**Hospital**), the number of patients on ventilation (**ICU_Beds**), the number of tests (**VirusTests**), the number of positive cases (**Cases**) and the number of deaths (**Deaths**). The first four data variables were gathered from the UK Government dashboard¹, while the last two variables were downloaded from the Johns Hopkins

¹Available at the website <https://coronavirus.data.gov.uk/>

University database². This information was available in cumulative form, therefore to obtain the daily time series the previous days total was subtracted from the current days total. As unstructured data, we collected Google Trends information on the number of searches for the keywords *Covid-19*, *coronavirus*, *first wave*, *second wave* and *variant*, using the `gtrendsR` package from the R software (Massicotte and Eddelbuettel, 2021; R Core Team, 2020). In addition, we retrieved Twitter messages containing the same keywords used to perform Google Trends searches, using the `rtweet` R package (Kearney, 2019). Three batches of 18,000 tweets were collected 3 times a day, everyday. We proceeded with data cleansing by removing tweets written from outside the UK and those produced from locations with less than 10 tweets. We also removed any duplicate tweets and those sent by automated accounts which contained factual information about daily case numbers or retweets of news stories. Finally, we removed tweets directly addressed to foreign political leaders or politicians, obtaining a final large Twitter dataset of 577,231 tweets. From the Twitter data, we considered the total number of tweets as well as the sentiment scores calculated using two different lexicons: Bing and Afinn (Hu and Liu, 2004), which are available in the R `tidytext` package (Silge and Robinson, 2016). The Bing lexicon splits words into positive or negative. The Bing sentiment score for each tweet is calculated by counting the number of positive words used in each tweet and subtracting from this the number of negative words. The Afinn lexicon scores words between ± 5 . The Afinn sentiment score is calculated by multiplying the score of each word by the number of times it appears in the tweet; these scores are then summed to derive the overall sentiment score.

Figure 1 shows the trace plots of the Covid-19 official and social media times series. The panels depict the variables (from top to bottom) `Admissions`, the Afinn sentiment scores (`Afinn`), the Bing sentiment scores (`Bing`), `Cases`, `Deaths`, the Google Trends searches (`Google`), `Hospital`, `ICU_Beds`, the total number of Tweets (`Tweets`) and `VirusTests`. We notice a higher volume of online messages produced at the beginning of the collection period and spikes corresponding to periods of more heated online discussions. In addition, the daily reporting of official structured data shows high variability and there is often a lag in reporting in the UK government figures due to figures being under reported at the weekend. The highlighted drawbacks of the official data could be overcome by data integration with social media information, which do not suffer from reporting lags.

²Available at the website <https://coronavirus.jhu.edu/>

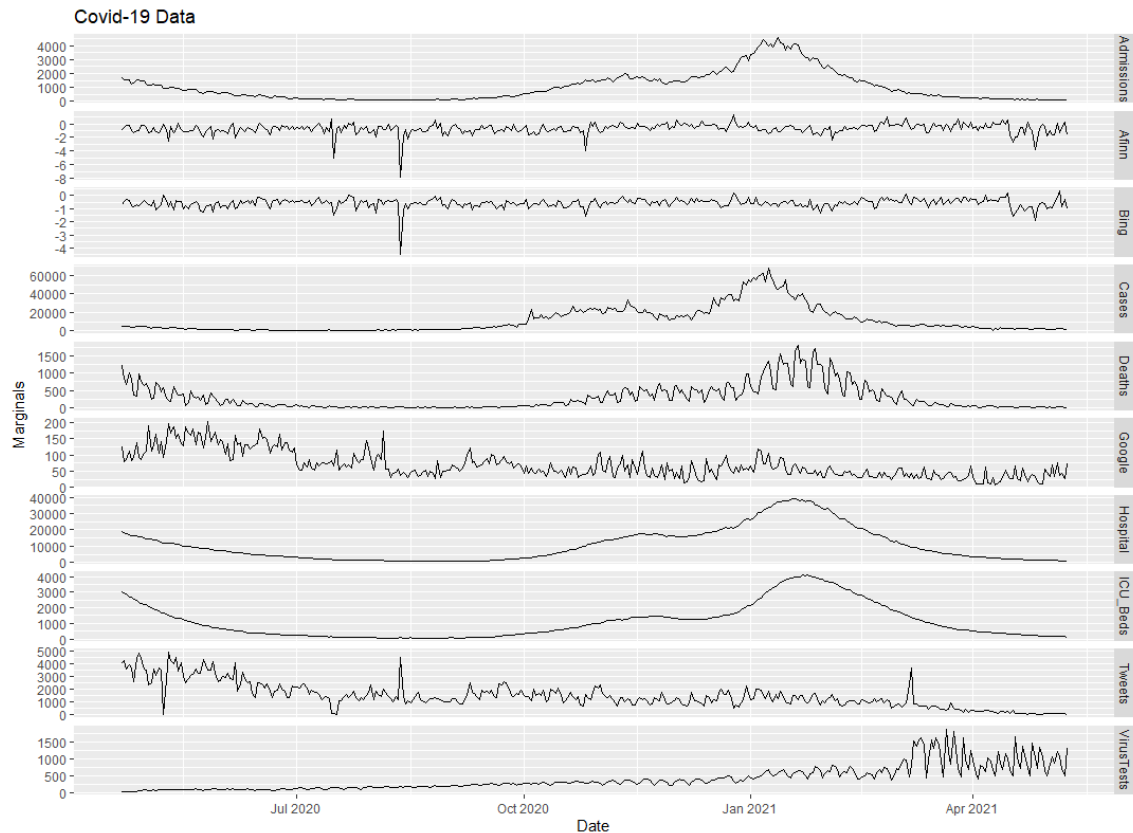


Figure 1: Trace plots of Covid-19 data.

3 Methodology

The copula is a function that allows us to bind together a set of marginals, to model their dependence structure and to obtain the joint multivariate distribution (Joe, 1997; Nelsen, 2007; Dalla Valle, 2017b,a). Sklar’s theorem (Sklar, 1959) is the most important result in copula theory. It states that, given a vector of random variables $\mathbf{X} = (X_1, \dots, X_d)$, with d -dimensional joint cumulative distribution function $F(x_1, \dots, x_d)$ and marginal cumulative distributions (cdf) $F_j(x_j)$, with $j = 1, \dots, d$, a d -dimensional copula C exists, such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d); \boldsymbol{\theta}),$$

where $F_j(x_j) = u_j$, with $u_j \in [0, 1]$ are called *u-data*, and $\boldsymbol{\theta}$ denotes the set of parameters of the copula. The joint density function can be derived as

$$f(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d); \boldsymbol{\theta}) \cdot f_1(x_1) \cdots f_d(x_d),$$

where c denotes the d -variate copula density. The copula allows us to determine the joint multivariate distribution and to describe the dependencies among the marginals, that can potentially be all different and can be modelled using distinct distributions.

In this paper, we adopt the 2-steps inference function for margins (IFM) approach (Joe and Xu, 1996), estimating the marginals in the first step, and then the copula, given the marginals, in the second step.

3.1 Marginal Models

Given the different characteristics of the ten marginals, we fitted different models for each of the ten time series. Further, we extracted the residuals ε_j , with $j = 1, \dots, d$, from each marginal model and we applied the relevant distribution functions to get the *u-data* $F_j(\varepsilon_j) = u_j$ to be plugged into the copula.

3.1.1 New admitted patients (Admissions)

The best fitting model for the **Admissions** marginal was the SHASHo2 model. This model belongs to the family of GAMLSS distributions, which stands for Generalised Additive Models for Location, Scale and Shape. GAMLSS are very flexible models, which include a wide range of continuous and discrete distributions (Stasinopoulos et al., 2017). The SHASHo2 model, developed by Jones and Pewsey (2009), is also known as Sinh-Arcsinh original type 2 distribution and depends on four parameters: μ the location parameter, σ the scaling parameter, ν the skewness parameter and τ the kurtosis parameter. The probability density function (pdf) of the SHASHo2 model is given by

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{c}{\sqrt{2\pi}} \frac{1}{\sqrt{1+z^2}} \exp\left(-\frac{r^2}{2}\right) \quad (1)$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$, where $z = (x - \mu) / (\sigma\tau^2)$, $r = \sinh[\tau \sinh^{-1}(z) - \nu]$ and $c = \cosh[\tau \sinh^{-1}(z) - \nu]$. We assumed that the parameter μ of the SHASHo2 model is related to time, as explanatory variable, through an appropriate link function, with coefficient β (for more details, see Rigby and Stasinopoulos (2005); Rigby et al. (2019)).

3.1.2 Afirm sentiment score (Afirm)

We fitted the **Afirm** marginal with a reparametrized version of Skew Student t type 3 model (SST), which, similarly to the previous marginal, belongs to the family of GAMLSS distributions and depends on four parameters: the mode (μ), scaling (σ), skewness (ν) and kurtosis (τ) (Fernández and Steel, 1998). The pdf for the SST model is

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left[1 + \frac{z^2}{\tau} \left(\nu^2 \mathbb{1}(x < \mu) + \frac{1}{\nu^2} \mathbb{1}(x \geq \nu) \right) \right]^{-\frac{\tau+1}{2}}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, where $z = (x - \mu)/\sigma$, $c = 2\nu[(1 + \nu^2)B(1/2, \tau/2)\tau^{1/2}]^{-1}$, B is the beta function and $\mathbb{1}(\cdot)$ is the indicator function. Similarly to the SHASHo2 model, for the SST model we assumed that the parameter μ is related to time, as explanatory variable, through an appropriate link function, with coefficient β .

3.1.3 Bing sentiment score (Bing)

The best model for **Bing** was the Normal-Exponential- t (NET) distribution. This is again a four parameter continuous distribution, belonging the GAMLSS family, which was introduced by Rigby and Stasinopoulos (1994). The parameters are: mean (μ), scaling (σ), first kurtosis parameter (ν) and second kurtosis parameter (τ). The pdf of the NET model is given by

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \begin{cases} \exp\left(-\frac{z^2}{2}\right) & \text{when } |z| \leq \nu \\ \exp\left(-\nu|z| + \frac{\nu^2}{2}\right) & \text{when } \nu < |z| \leq \tau \\ \exp\left(-\nu\tau \log\left(\frac{|z|}{\tau}\right) - \nu\tau + \frac{\nu^2}{2}\right) & \text{when } |z| > \tau \end{cases}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > \max(\nu, \nu^{-1})$, where $z = (x - \mu)/\sigma$ and $c = (c_1 + c_2 + c_3)^{-1}$, $c_1 = \sqrt{2\pi}[2\Phi(\nu) - 1]$, $c_2 = \frac{2}{\nu} \exp\left(-\frac{\nu^2}{2}\right)$ and $c_3 = \frac{2}{(\nu\tau-1)\nu} \exp\left(-\nu\tau + \frac{\nu^2}{2}\right)$, with Φ the cumulative distribution function of the standard normal distribution. As with the previous marginals, we assumed that the parameter μ of the NET model is related to time.

3.1.4 Number of positive cases (Cases)

We fitted the **Cases** marginal with an ARIMA-GARCH model with Student's t innovations. This model combines the features of the autoregressive integrated

moving average (ARIMA) model with the generalized autoregressive conditional heteroskedastic (GARCH) model, allowing us to capture time series volatility over time (for more information see, for example Hyndman and Athanasopoulos (2018)). The GARCH model is typically denoted as GARCH(p, q), with parameters p and q , where p is the number of lag residuals errors and q is the number of lag variances. The ARIMA(p, d, q)-GARCH(p, q) model can be expressed as:

$$y_t = a + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

$$\varepsilon_t = \sqrt{\sigma_t} z_t \quad \sigma^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^q \beta_i \sigma_{t-i}^2 \quad (2)$$

where the first line is the ARIMA part of the model, while the second line is the GARCH part of the model. Also, $y_t = (1 - B)^d x_t$, x_t are the original data values, B is the backshift operator, a is a constant, ϕ_i are the autoregressive parameters, θ_i are the moving average parameters; α_i and β_i are the parameters of the GARCH part of the model, and ε_t follows a Student's t distribution.

3.1.5 Number of deaths (Deaths)

The best model for the **Deaths** marginal was the SHASHo model, whose acronym stands for Original Sinh-Arcsinh distribution. This model is very similar to the SHASHo2, represented in Eq.(1). The pdf of the SHASHo model is

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \frac{\tau}{\sqrt{2\pi}} \frac{1}{\sqrt{1+z^2}} \exp\left(-\frac{r^2}{2}\right) \quad (3)$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$, where $z = (x - \mu)/(\sigma\tau)$, $r = \sinh[\tau \sinh^{-1}(z) - \nu]$ and $c = \cosh[\tau \sinh^{-1}(z) - \nu]$. As for the other marginals fitted with GAMLSS-type models, we assumed that the parameter μ of the SHASHo depends on time.

3.1.6 Google trends (Google)

Since **Google** includes values equal to zero, we fitted a Tweedie Generalised Linear Model for this marginal (Dunn and Smyth, 2018). The Tweedie distribution has nonnegative support and can have a discrete mass at zero, making it useful to model responses that are a mixture of zeros and positive values. The Tweedie distribution

belongs to the exponential family, assuming the following mean and variance, for $t = 1, \dots, T$:

$$E[x_t] = \mu_t^q = \zeta_t b, \quad \text{Var}[x_t] = \phi \mu_t^p$$

where ζ_t denotes the time covariate, b is the associated regression coefficient, ϕ is the dispersion parameter and p and q are extra parameters that control the mean and variance of the distribution, respectively.

3.1.7 Number of hospital cases (Hospital)

The best model for the `Hospital` marginal was the ARIMA-GARCH model with Student's t innovations, as illustrated in Eq.(2).

3.1.8 Number of patients on ventilation (ICU_Beds)

The best fitting model for the `ICU_Beds` marginal was the SHASHo model, with pdf given by Eq.(3).

3.1.9 Total number of tweets (Tweets)

We fitted the `Tweets` marginal with a Skew Exponential Power type 4 (SEP4) model, which is a four parameter distribution belonging to the GAMLSS family (Rigby and Stasinopoulos, 2005). This is a “spliced-shaped” distribution with the following pdf

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} [\exp(-|z|^\nu) \mathbb{1}(x < \mu) + \exp(-|z|^\tau) \mathbb{1}(x \geq \mu)]$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, where $z = (x - \mu)/\sigma$, $c = (\Gamma(1 + \nu^{-1}) + \Gamma(1 + \tau^{-1}))^{-1}$, Γ is the gamma function and $\mathbb{1}(\cdot)$ is the indicator function. Note that μ is the mode of Y . Here we assumed that the parameter μ is related to time, as explanatory variable.

3.1.10 Number of tests (VirusTests)

The best model for the `VirusTests` marginal was the ARIMA-GARCH model with Student's t innovations, as illustrated in Eq.(2), fitted on the the number of test adjusted by 1000.

3.2 Vine Copula Model

A *vine copula* (or *vine*) represents the pattern of dependence of multivariate data via a cascade of bivariate copulas, allowing us to construct flexible high-dimensional copulas using only bivariate copulas as building blocks. For more details about vine copulas see Czado (2019).

In order to obtain a vine copula we proceed as follows. First we factorise the joint distribution $f(x_1, \dots, x_d)$ of the random vector $\mathbf{X} = (X_1, \dots, X_d)$ as a product of conditional densities

$$f(x_1, \dots, x_d) = f_d(x_d) \cdot f_{d-1|d}(x_{d-1}|x_d) \cdot \dots \cdot f_{1|2\dots d}(x_1|x_2, \dots, x_d). \quad (4)$$

The factorisation in (4) is unique up to re-labelling of the variables and it can be expressed in terms of a product of bivariate copulas. In fact, by Sklar's theorem, the conditional density of $X_{d-1}|X_d$ can be easily written as

$$f_{d-1|d}(x_{d-1}|x_d) = c_{d-1,d}(F_{d-1}(x_{d-1}), F_d(x_d); \boldsymbol{\theta}_{d-1,d}) \cdot f_{d-1}(x_{d-1}), \quad (5)$$

where $c_{d-1,d}$ is a bivariate copula, with parameter vector $\boldsymbol{\theta}_{d-1,d}$. Through a straightforward generalisation of Eq.(5), each term in (4) can be decomposed into the appropriate bivariate copula times a conditional marginal density. More precisely, for a generic element X_j of the vector \mathbf{X} we obtain

$$f_{X_j|\mathbf{v}}(x_j|\mathbf{v}) = c_{X_j, \nu_\ell; \mathbf{v}_{-\ell}}(F_{X_j|\mathbf{v}_{-\ell}}(x_j|\mathbf{v}_{-\ell}), F_{\nu_\ell|\mathbf{v}_{-\ell}}(\nu_\ell|\mathbf{v}_{-\ell}); \boldsymbol{\theta}_{X_j, \nu_\ell; \mathbf{v}_{-\ell}}) \cdot f_{X_j|\mathbf{v}_{-\ell}}(x_j|\mathbf{v}_{-\ell}), \quad (6)$$

where \mathbf{v} is the conditioning vector, ν_ℓ is a generic component of \mathbf{v} , $\mathbf{v}_{-\ell}$ is the vector \mathbf{v} without the component ν_ℓ , $F_{X_j|\mathbf{v}_{-\ell}}(\cdot|\cdot)$ is the conditional distribution of X_j given $\mathbf{v}_{-\ell}$ and $c_{X_j, \nu_\ell; \mathbf{v}_{-\ell}}(\cdot, \cdot)$ is the conditional bivariate copula density, which can typically belong to any family (e.g. Gaussian, Student's t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8, etc.; for more information on copula families, see Nelsen (2007)), with parameter $\boldsymbol{\theta}_{X_j, \nu_\ell; \mathbf{v}_{-\ell}}$. The d -dimensional joint multivariate distribution function can hence be expressed as a product of bivariate copulas and marginal distributions by recursively plugging Eq.(6) in Eq.(4).

For example, let us consider a 6-dimensional distribution. Then, Eq.(4) translates to

$$f(x_1, \dots, x_6) = f_6(x_6) \cdot f_{5|6}(x_5|x_6) \cdot f_{4|5,6}(x_4|x_5, x_6) \cdot \dots \cdot f_{1|2\dots 6}(x_1|x_2, \dots, x_6). \quad (7)$$

The second factor $f_{5|6}(x_5|x_6)$ on the right-hand side of (7) can be easily decomposed into the bivariate copula $c_{5,6}(F_5(x_5), F_6(x_6))$ and marginal density $f_5(x_5)$:

$$f_{5|6}(x_5|x_6) = c_{5,6}(F_5(x_5), F_6(x_6); \boldsymbol{\theta}_{5,6}) \cdot f_5(x_5).$$

On the other hand, the third factor on the right-hand side of (7) can be decomposed using the (6) as

$$f_{4|5,6}(x_4|x_5, x_6) = c_{4,5;6}(F_{4|6}(x_4|x_6), F_{5|6}(x_5|x_6); \boldsymbol{\theta}_{4,5;6}) \cdot f_{4|6}(x_4|x_6).$$

Therefore, one of the possible decompositions of the joint density $f(x_1, \dots, x_6)$ is given by the following expression, which includes the product of marginal densities and copulas, which are all bivariate:

$$f(x_1, \dots, x_6) = \prod_{j=1}^6 f_j(x_j) \cdot c_{1,2} \cdot c_{1,3} \cdot c_{3,4} \cdot c_{1,5} \cdot c_{5,6} \cdot c_{2,3;1} \cdot c_{1,4;3} \cdot c_{3,5;1} \cdot c_{1,6;5} \\ \cdot c_{2,4;1,3} \cdot c_{4,5;1,3} \cdot c_{3,6;1,5} \cdot c_{2,5;1,3,4} \cdot c_{4,6;1,3,5} \cdot c_{2,6;1,3,4,5}. \quad (8)$$

Eq.(8) is called *pair copula construction*. Note that in the previous equation the notation has been simplified, setting $c_{a,b} = c_{a,b}(F_a(x_a), F_b(x_b); \boldsymbol{\theta}_{a,b})$.

Two particular types of vines are the Gaussian vine and the Independence vine. The first one is constructed using solely Gaussian bivariate pair-copulas as building blocks, such that each conditional bivariate copula density $c_{X_J, \nu_\ell; \mathbf{V}_{-\ell}}(\cdot, \cdot)$ described in Eq.(6) is a Gaussian copula. The second type is the independence vine, which is constructed using only independence pair-copulas, that are simply given by the product of the marginal distributions of the random variables. In this latter case each conditional bivariate copula density $c_{X_J, \nu_\ell; \mathbf{V}_{-\ell}}(\cdot, \cdot)$ described in Eq.(6) is an Independence copula, implying absence of dependence between the variables.

Pair copula constructions can be represented through a graphical model called *regular vine* (R-vine). An R-vine $\mathcal{V}(d)$ on d variables is a nested set of trees (connected acyclic graphs) T_1, \dots, T_{d-1} , where the variables are represented by nodes linked by edges, each associated with a certain bivariate copula in the corresponding pair copula construction. The edges of tree T_k are the nodes of tree T_{k+1} , $k = 1, \dots, d-1$. Two edges can share a node in tree T_k without the associated nodes in tree T_{k+1} being connected. In an R vine, two edges in T_k which become two nodes in tree T_{k+1} , can only share an edge if in tree T_k the edges shared a common node, but they are not necessarily connected by an edge. Figure 2 shows the 6-dimensional R-vine represented in Eq.(8). Each edge corresponds to a pair copula density (possibly belonging to different families) and the edge label corresponds to the subscript of the pair copula density, e.g. edge 2, 4; 1, 3 corresponds to the copula $c_{2,4;1,3}$.

In order to estimate the vine, its structure as well as the copula parameters have to be specified. A sequential approach is generally adopted to select a suitable R-vine decomposition, specifying the first tree and then proceeding similarly for the

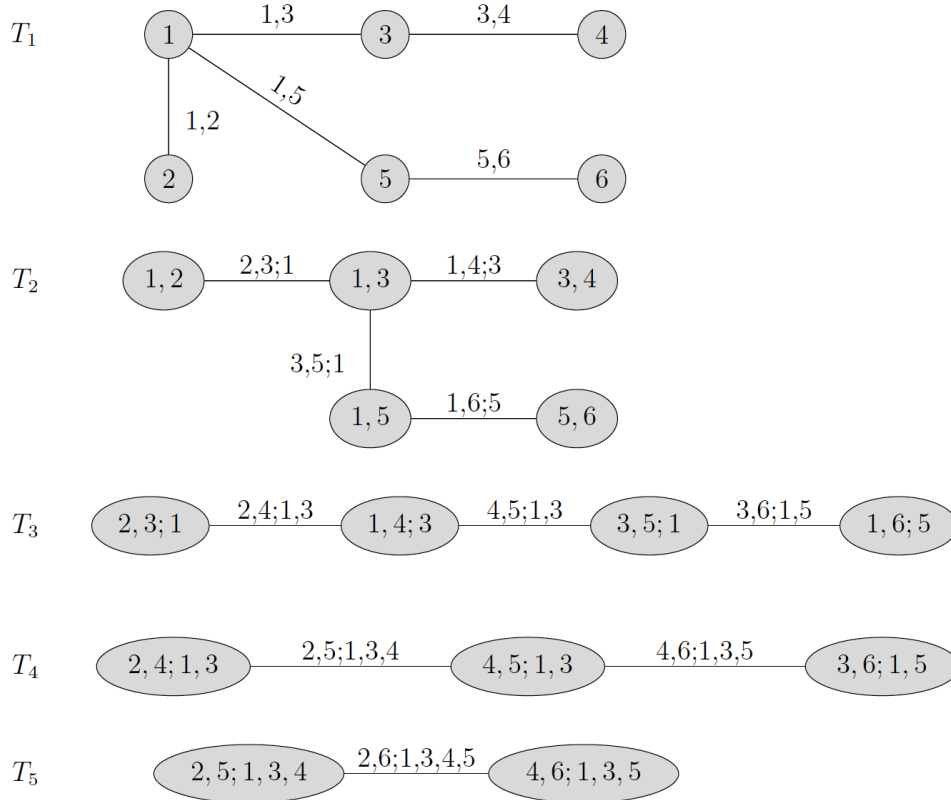


Figure 2: Six-dimensional R-vine graphical representation. *Source: Czado (2019)*

following trees. For selecting the structure of each tree, we followed the approach suggested by Aas et al. (2009) and developed by Dissmann et al. (2013), using the maximal spanning tree algorithm. This algorithm defines a tree on all nodes (named spanning tree), which maximizes the sum of absolute pairwise dependencies, measured, for example, by Kendall's τ . This specification allows us to capture the strongest dependencies in the first tree and to obtain a more parsimonious model. Given the selected tree structure, a copula family for each pair of variables is identified using the Akaike Information Criterion (AIC), or the Bayesian Information Criterion (BIC). This choice is typically made amongst a large set of families, comprising elliptical copulas (Gaussian and Student's t) as well as Archimedean copulas (Clayton, Gumbel, Frank and Joe), their mixtures (BB1, BB6, BB7 and BB8) and their rotated versions, to cover a large range of possible dependence structures. For an overview of the different copula families, see Joe (1997) or Nelsen

(2007). The copula parameters θ for each pair-copula in the vine are estimated using the maximum likelihood (MLE) method, as illustrated by Aas et al. (2009). The R-vine estimation procedure is repeated for all the trees, until the R-vine is completely specified.

3.3 Out-of-sample predictions

In order to evaluate the suitability of the proposed vine copula model in relation to other methods, we produced one-day-ahead out-of-sample predictions and we compared them to the original data. Let $\mathbf{X} = \{\mathbf{X}_t; t = 1, \dots, T\}$ be the 10-dimensional time series of Covid-19 and social media data. Our aim is to forecast \mathbf{X}_{T+1} based on the information available at time T . In order to do that, we adopted the forecasting method described by Simard and Rémillard (2015). Before fitting the vine, we extracted the residuals from the marginals, as explained in Section 3.1, and obtained the *u-data*. Next, after fitting the vine, we simulated M realizations from the vine copula. Hence, we calculated the predicted values for each simulation, using the inverse cdf and the relevant fitted marginal models. More precisely, we applied the inverse transformation to the M realizations from the vine copula to obtain the residuals which we then plugged into the marginal models to get the predicted values of the official variables. Then, we calculated the average prediction for all simulations $\hat{\mathbf{X}}_{T+1}^{Avg}$ and use it to forecast \mathbf{X}_{T+1} . The prediction interval of level $(1 - \alpha) \in (0, 1)$ for \mathbf{X}_{T+1} was calculated by taking the estimated quantiles of order $\alpha/2$ and $1 - \alpha/2$ amongst the simulated data. We denote by $\hat{\mathbf{X}}_{T+1}^l$ and $\hat{\mathbf{X}}_{T+1}^u$ the lower and upper values of the prediction intervals.

In order to compare and contrast the accuracy of predictions for different models, we made use of two indicators: the mean squared error (MSE) to evaluate point forecasts and the mean interval score (MIS), proposed by Gneiting and Raftery (2007), to assess the accuracy of the prediction intervals. The MSE for each variable $j = 1, \dots, d$ was calculated as follows

$$\text{MSE}_j = \frac{1}{S} \sum_{t=T+1}^{T+S} (x_{t,j} - \hat{x}_{t,j})^2$$

where $x_{t,j}$ is the observed value for each variable at each time point t , $\hat{x}_{t,j}$ is the corresponding predicted value, $T + 1$ denotes the first predicted date, while $T + S$ indicates the last predicted date. The 95% MIS for each variable, at level $\alpha = 0.05$,

was computed as

$$\text{MIS}_j = \frac{1}{S} \sum_{t=T+1}^{T+S} \left[(\hat{x}_{t,j}^u - \hat{x}_{t,j}^l) + \frac{2}{\alpha} (\hat{x}_{t,j}^l - x_{t,j}) \mathbb{1}(x_{t,j} < \hat{x}_{t,j}^l) + \frac{2}{\alpha} (x_{t,j} - \hat{x}_{t,j}^u) \mathbb{1}(x_{t,j} > \hat{x}_{t,j}^u) \right]$$

where $\hat{x}_{t,j}^l$ and $\hat{x}_{t,j}^u$ denote, respectively, the lower and upper limits of the prediction intervals for each variable at each time point, and $\mathbb{1}(\cdot)$ is the indicator function.

4 Result Analysis and Discussion

We now present the results of the analysis of the official and online-retrieved Covid-19 data.

4.1 Twitter Wordclouds

First, we analysed the information gathered on Twitter, cleaning and stemming the tweets and producing graphical representations of the data using wordclouds.

Figure 3 displays the wordcloud obtained from the collected tweets discussing Covid-19 in the UK. The most frequent words are related to “people” and the effects of the pandemic on them. We can also notice the names of the most prominent politicians and words related to political decisions.

Figure 4 shows the sentiment wordcloud created from the collected tweets, obtained with the Bing method. This data visualization highlights the positive words in blue and the negative words in pink. The most popular positive words are related to the “support” received throughout the pandemic, while the most popular negative words are related to the worst consequences of Covid-19 on the health of individuals.

4.2 Marginals Estimation

Table 1 lists the parameter estimates, obtained via the MLE method, of the best fitting models for the marginals, as described in Section 3.1. Standard errors are in brackets.

As an example, Figure 5 shows the fit of the residuals for the **Tweets** marginal. The top panel displays the QQ-plot comparing the Gaussian theoretical quantiles with the sample quantiles, the middle panel illustrates the observations (black line) and in-sample predictions obtained from the fitted SEP4 model (red line), while the bottom panel shows the histogram of the resulting *u-data*. The plots clearly show an excellent fit of the SEP4 model to the marginal, as demonstrated by the points in the

Table 1: Parameter estimates of the marginals. Standard errors are in brackets.

| Marginal | Parameter | Estimate |
|---------------------------------------|------------|------------------------|
| Admissions SHASHo2 | ν | 2.5812 (0.0369) |
| | σ | 3.1447 (0.0588) |
| | μ | -710.0801 (1103.0128) |
| | τ | -0.4575 (0.0178) |
| Afinn SST | ν | -0.2901 (0.0635) |
| | σ | -0.3719 (0.0441) |
| | μ | -23.4129 (4.356) |
| | τ | 0.6298 (0.2170) |
| Bing NET | ν | 1.5 (fixed) |
| | σ | -1.4358 (0.0658) |
| | μ | -6.3868 (0.0145) |
| | τ | 2 (fixed) |
| Cases ARIMA(1,0,2)-GARCH(1,1) | a | 11231.83 (0.0042) |
| | ϕ_1 | 1.0000 (0.0020) |
| | θ_1 | -0.1968 (0.0532) |
| | θ_2 | -0.1144 (0.0466) |
| | ω | 193779.5 (0.0004) |
| | α_1 | 0.4764 (0.0705) |
| | β_1 | 0.5226 (0.0793) |
| | σ | 2.9812 (0.3571) |
| Deaths SHASHo | ν | 2.5756 (0.0363) |
| | σ | 0.4274 (0.0687) |
| | μ | -45.8608 (119.0992) |
| | τ | -0.6352 (0.0142) |
| Google Tweedie | b | -0.0037 (0.0002) |
| | ϕ | 0.0402 (fixed) |
| Hospital ARIMA(1,0,1)-GARCH(2,1) | a | 19141.45(0.0188) |
| | ϕ_1 | 1.0000. (0.0010) |
| | θ_1 | 0.5473 (0.0306) |
| | ω | 103266.1 (0.0002) |
| | α_1 | 0.0156 (0.0777) |
| | α_2 | 0.9834 (0.2979) |
| | β_1 | 0.0000 (0.0646) |
| | σ | 2.4623 (0.0735) |
| ICU_Beds SHASHo | ν | 2.7742 (0.0385) |
| | σ | 2.3202 (0.0010) |
| | μ | -6700.6665 (0.0000094) |
| | τ | -0.4783 (0.0142) |
| Tweets SEP4 | ν | 0.1235 (0.1098) |
| | σ | 6.1133 (0.1428) |
| | μ | 120943.3 (4.765) |
| | τ | -0.1197 (0.0887) |
| VirusTests ARIMA(2,0,1)-GARCH(1,1) | a | 49.8691 (6.6710) |
| | ϕ_1 | 1.5807 (0.0009) |
| | ϕ_2 | -0.5766 (0.0007) |
| | θ_1 | -0.9412 (0.0169) |
| | ω | 28.3309 (1.1299) |
| | α_1 | 0.3309 (0.0464) |
| | β_1 | 0.6681 (0.0414) |
| | σ | 14.5705 (9.6972) |

Wordcloud for Tweets about Covid-19 in the UK

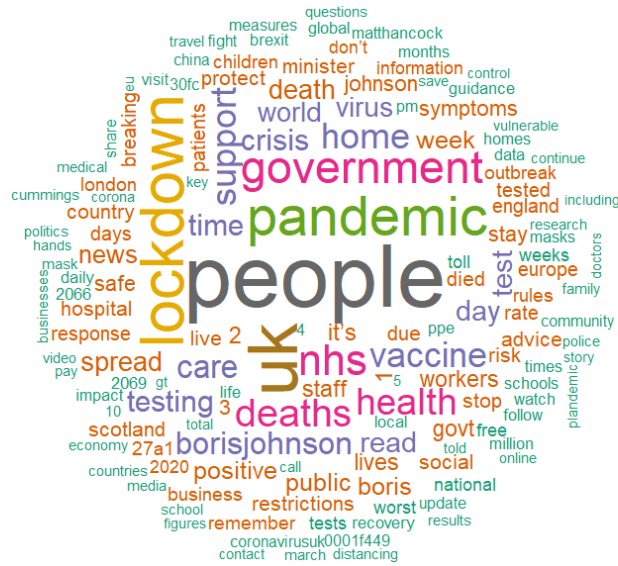


Figure 3: Wordcloud of the UK Covid-19 data.

QQ-plot aligning well to the main diagonal, the in-sample predictions overlapping the observed data and the shape of the *u*-data histogram being close to a uniform pattern.

4.3 Vine Estimation

Once the marginals were estimated, we derived the corresponding *u*-data from the residuals, as illustrated in Section 3.1. Then, we carried out fitting and model selection for the vine copula for each location using the R package `VineCopula` (Nagler et al., 2021).

Figure 6 displays the first tree of the vine copula for the Covid-19 data. The nodes are denoted with blue squares, with the numbers corresponding to the margins reported on them. On each edge, the plot shows the name of the selected pair copula family and the estimated copula parameter expressed as Kendall's τ . In order to estimate the vines, we adopted the Kendall's τ criterion for tree selection, the AIC

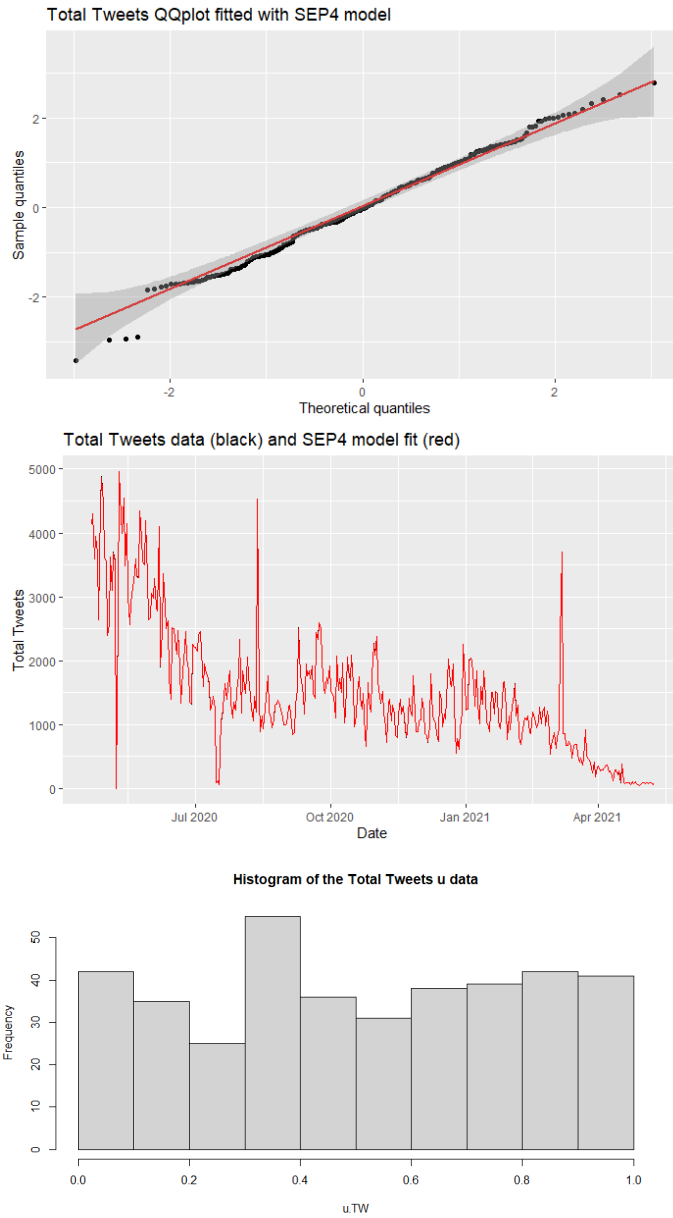


Figure 5: Plot illustrating the fit of the residuals for the `Tweets` marginal. Top plot: QQ-plot comparing the Gaussian theoretical quantiles with sample quantiles. Middle plot: observed time series (black line) and in-sample predictions obtained from the fitted SEP4 model (red line). Bottom plot: Histogram of the resulting *u-data*.

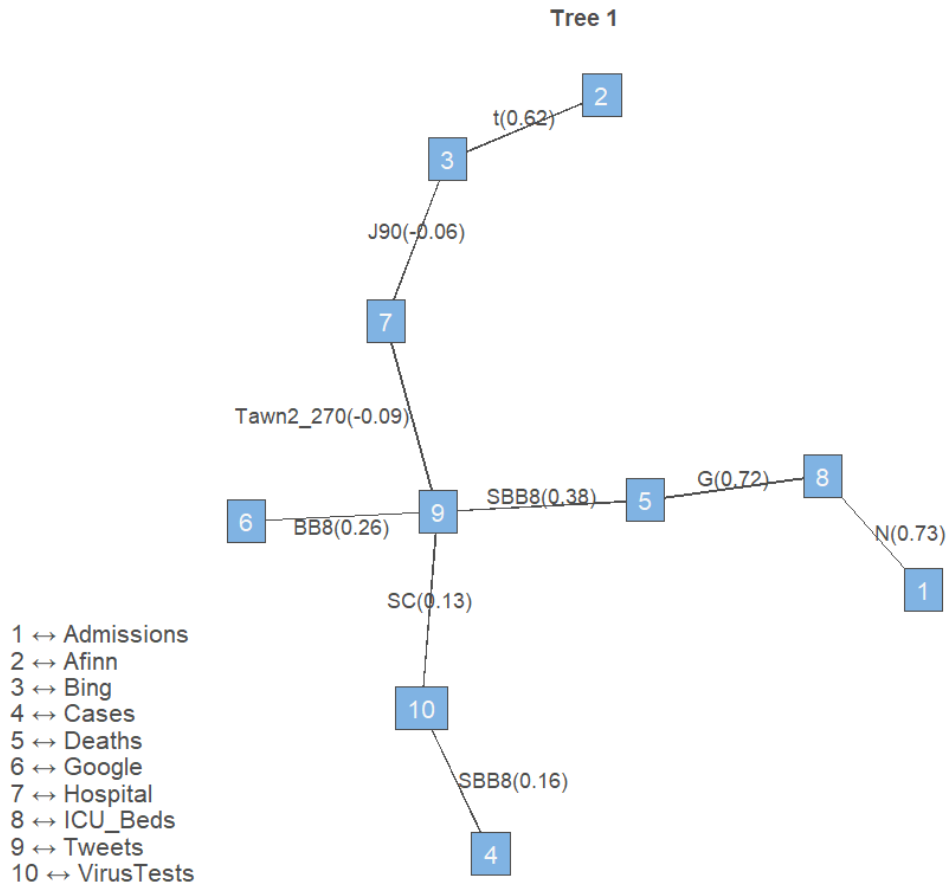


Figure 6: First tree of the vine copula for the Covid-19 data. The legend shows the names of the variables displayed on each node. The pair-copula families are shown on the edges and the Kendall's τ s are given in brackets.

copulas such as the BB8 (Joe-Frank), that can accommodate various dependence shapes. Most of the associations between the variables are positive. The strongest associations are between the official Covid-19 variables `ICU_Beds` and `Admissions`, between `ICU_Beds` and `Deaths` and between the `Bing` and `Afinn` sentiment scores. Also, `Deaths` and `Tweets` are mildly associated.

4.4 Out-of-sample prediction results

In this Section we constructed out-of-sample predictions using the proposed vine methodology, which integrates official and social media Covid-19 variables. We then compared the predictions obtained with our methodology with those yielded using two traditional approaches. The former is based on vines built exclusively using Gaussian pair copulas, which are the most common in applications, but are restricted to dependence symmetry and absence of tail dependence. The latter approach assumes independence among the ten time series under consideration and therefore calculates predictions ignoring any association between official and online information.

Out-of-sample predictions based on the proposed model were constructed as illustrated in Section 3.3, considering the vine copula estimated as explained in Section 4.3 until the 1st April 2021 and using it to predict the period between the 2nd April 2021 and the 9th May 2021.

Table 2: MSEs calculated for each variable. The figures show the vine copula results (second column), the results assuming all Gaussian pair-copulas (third column), and assuming independence among variables (fourth column). The MSEs of the best performing approaches for each variable are in boldface.

| Marginal | Vine Copula | Gaussian | Independent |
|-------------------------|-----------------|-----------------|-----------------|
| <code>Admissions</code> | 12510.57 | 12515.63 | 12506.15 |
| <code>Afinn</code> | 1.0226 | 1.0667 | 0.9766 |
| <code>Bing</code> | 0.2417 | 0.2597 | 0.2452 |
| <code>Cases</code> | 577363 | 577332.3 | 577401.1 |
| <code>Deaths</code> | 886.7999 | 890.7506 | 887.5182 |
| <code>Google</code> | 385.3506 | 382.9412 | 384.9532 |
| <code>Hospital</code> | 1348411 | 1348573 | 1348496 |
| <code>ICU_Beds</code> | 49150.17 | 49142.98 | 49144.45 |
| <code>Tweets</code> | 17593.4 | 17583.2 | 17585.62 |
| <code>VirusTests</code> | 292876 | 292943.3 | 292902.5 |

Table 3: MISs calculated for each variable. The figures show the vine copula results (second column), the results assuming all Gaussian pair-copulas (third column), and assuming independence among variables (fourth column). The MISs of the best performing approaches for each variable are in boldface.

| Marginal | Vine Copula | Gaussian | Independent |
|------------|-----------------|-----------------|----------------|
| Admissions | 25.5155 | 25.5208 | 25.5101 |
| Afinn | 0.2510 | 0.2587 | 0.2461 |
| Bing | 0.1187 | 0.1231 | 0.1148 |
| Cases | 173.1329 | 173.1258 | 173.1414 |
| Deaths | 6.5837 | 6.5991 | 6.5867 |
| Google | 4.5348 | 4.1436 | 4.1482 |
| Hospital | 257.4757 | 257.4917 | 257.4841 |
| ICU_Beds | 49.3001 | 49.2965 | 49.2972 |
| Tweets | 28.9344 | 28.9253 | 28.9278 |
| VirusTests | 120.5653 | 120.5799 | 120.5709 |

Tables 2 and 3 list the MSE and MIS values calculated for each variable. The second columns show the vine copula results, the third columns show the results assuming all Gaussian pair-copulas and the fourth columns show the results assuming independence among variables. The MSEs and MISs of the best performing approaches for each variable are highlighted in boldface. Tables 2 and 3, show a similar model performance across the ten variables. According to both the MSE and MIS indicators, the vine copula approach outperforms the other two approaches for predicting the variables `Deaths`, `Hospital` and `VirusTest`. The Gaussian vine approach also performs well with several variables, while the independent vine approach seems to exceed the other two approaches only with the variables `Admissions` and `Afinn`.

The official Covid-19 variables `Admissions`, `Cases`, `Deaths`, `Hospital`, `ICU_Beds` and `VirusTests` are generally better predicted by the vine method, as opposed to the Gaussian and independence methods. This last approach assumes no dependence between any of the variables involved in the model. Hence, this approach indicates the absence of any association between the official and the social media variables, implying the lack of contribution of online-generated information in predicting the official Covid-19 variables. On the contrary, the vine approach assumes the presence of a dependence structure between the variables and, in particular, between the official and social media insights. Therefore, the better performance of the vine compared to the independence model demonstrates usefulness of social media

information in forecasting official Covid-19 variables.

The prediction of online-generated information (**Afinn**, **Bing**, **Google** and **Tweets**) also benefits from data integration. Indeed, most of the social media variables are more accurately forecasted by the vine model, particularly the Gaussian one. This indicates that the Gaussian approach, characterized by a symmetric dependence structure, is flexible enough to model the social media variables.

Figure 7 shows the forecasts and prediction intervals for the the official Covid-19 variables **Admissions**, **Cases** (first row), **Deaths**, **Hospital** (second row), **ICU_Beds** and **VirusTests** (third row), obtained with the vine copula methodology for the period between the 2nd April 2021 and the 9th May 2021. The black lines denote the observed values, the inner red lines denote the predicted values and the outer dotted red lines denote the 95% prediction intervals. We notice that intervals predicted by the vine copula method capture most of the dynamic of the official Covid-19 variables, indicating that the proposed methodology is able to leverage social media information for forecasting official Covid-19-related data.

5 Concluding Remarks

In this paper, we propose a new methodology aimed at obtaining more accurate forecasts compared to traditional approaches, for variables measuring the Covid-19 dynamics. The proposed methodology is based on the integration of Covid-19 variables collected from official UK sources with online generated social media insights, relevant to the same geographical area. Together with official Covid-19 information related to infection counts and the pressure on the national health service, we also gathered Google Trends searches and Twitter microblogging messages involving keywords related to the Covid-19 pandemic. From the tweets, we considered the volume as well as the sentiment scores, to investigate the feelings of people towards the pandemic. Our methodology is based on vine copulas, which are able to model the dependence structure between the marginals, and thus to take advantage of the association between official Covid-19 and social media variables. We tested our approach calculating out-of-sample predictions and comparing the vine copula method with two traditional approaches: the first based on a vine constructed with all Gaussian copulas, and the second based on independence between variables. The results show that the vine copula method outperforms the other two approaches for predicting the number of deaths, hospital admissions and tests, demonstrating that our methodology is able to leverage social media information to obtain more accurate predictions of Covid-19 effects than the other two approaches. In some cases, the Gaussian vine copula method is selected, showing that the vine data integration

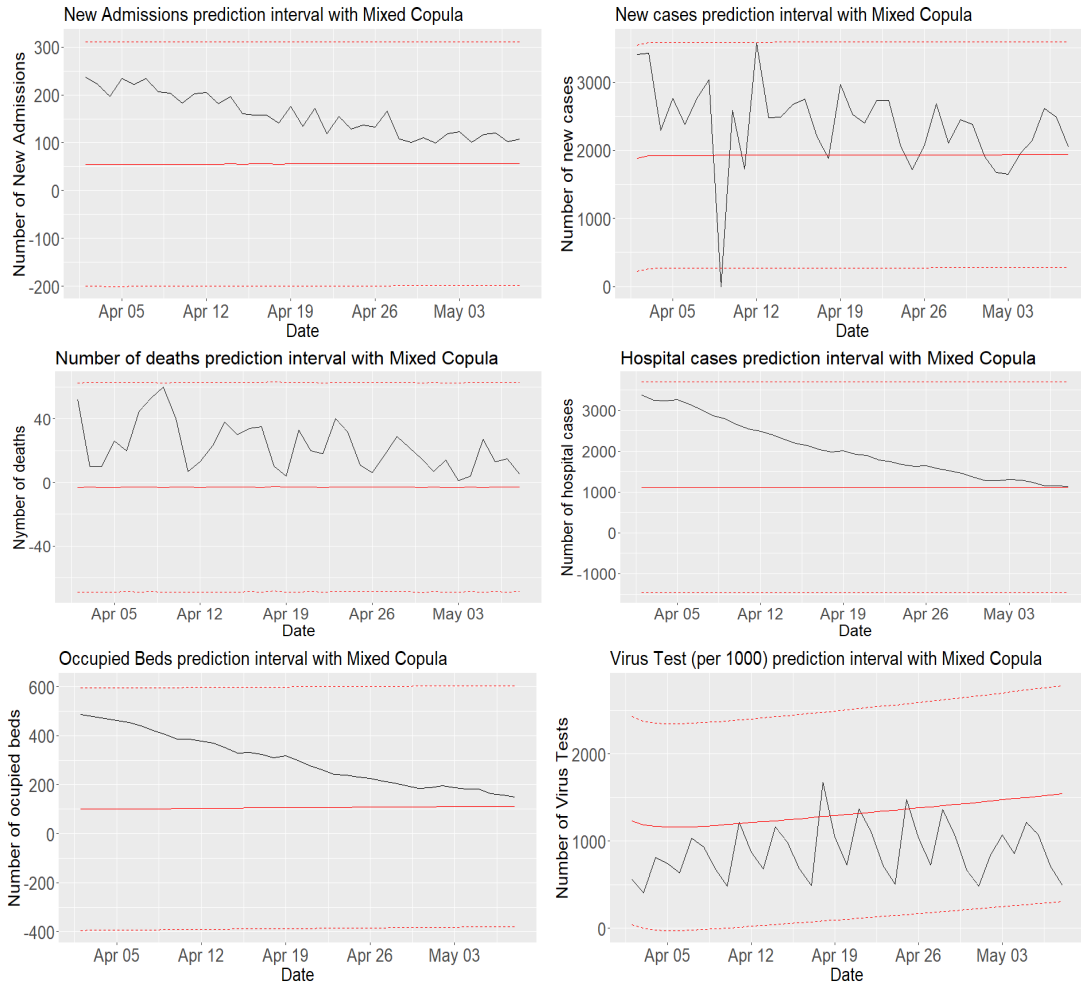


Figure 7: Line plots showing forecasts and prediction intervals for the official Covid-19 variables **Admissions**, **Cases** (first row), **Deaths**, **Hospital** (second row), **ICU_Beds** and **VirusTests** (third row), obtained with the vine copula methodology for the period between the 2nd April 2021 and the 9th May 2021. Observed values are in black, predicted values are the inner red lines and 95% prediction intervals are the outer red lines.

approach is still achieving the best performance, although some variables are less affected by asymmetries and tail dependence.

The proposed methodology will support policy makers to understand, monitor and combat the pandemic, assisting key medical and governmental actors to make informed decisions and to efficiently and effectively plan and allocate necessary resources.

Further investigations including additional social media information will be the object of future work. Also, the proposed approach could be extended using a 7 day rolling average in the model to adjust for the delay due to UK government figures being under reported at the weekend. Another extension will involve Bayesian inference, which would allow us to incorporate other information, such as experts' opinion, in the model. In addition, the use of more sophisticated machine learning approaches could be envisaged for deriving the sentiment variables to improve the proposed methodology.

Acknowledgements

This work was supported by the European Regional Development Fund project *Environmental Futures & Big Data Impact Lab*, funded by the European Structural and Investment Funds, grant number 16R16P01302 .

References

- Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics* 44(2), 182–198.
- Ahmed, I., M. Ahmad, G. Jeon, and F. Piccialli (2021). A framework for pandemic prediction using big data analytics. *Big Data Research* 25, 100190.
- Ansell, L. and L. Dalla Valle (2021). Social media integration of flood data: A vine copula-based approach. *arXiv preprint arXiv:2104.01869*.
- Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*.
- Dalla Valle, L. (2014). Official statistics data integration using copulas. *Quality Technology & Quantitative Management* 11(1), 111–131.

- Dalla Valle, L. (2017a). Copula and vine modeling for finance. In *Wiley StatsRef: statistics reference online*.
- Dalla Valle, L. (2017b). Copulas and vines. In *Wiley StatsRef: statistics reference online*.
- Dalla Valle, L. (2017c). Data integration. In *Wiley StatsRef: statistics reference online*.
- Dalla Valle, L. and R. Kenett (2018). Social media big data integration: A new approach based on calibration. *Expert Systems with Applications* 111, 76–90.
- Dalla Valle, L. and R. S. Kenett (2015). Official statistics data integration for enhanced information quality. *Quality and Reliability Engineering International* 31(7), 1281–1300.
- DeCaprio, D., J. Gartner, T. Burgess, K. Garcia, S. Kothari, S. Sayed, and C. J. McCall (2020). Building a covid-19 vulnerability index. *arXiv preprint arXiv:2003.07347*.
- Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis* 59, 52–69.
- Dunn, P. K. and G. K. Smyth (2018). *Generalized linear models with examples in R*. Springer.
- Fernández, C. and M. F. Steel (1998). On bayesian modeling of fat tails and skewness. *Journal of the american statistical association* 93(441), 359–371.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102(477), 359–378.
- Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177.
- Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.
- Jewell, N. P., J. A. Lewnard, and B. L. Jewell (2020). Predictive mathematical models of the covid-19 pandemic: underlying principles and value of projections. *Jama* 323(19), 1893–1894.

- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.
- Joe, H. and J. J. Xu (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia.
- Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. *Biometrika* *96*(4), 761–780.
- Kearney, M. W. (2019). rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software* *4*(42), 1829. R package version 0.7.0.
- Li, C., L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen (2020). Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020. *Eurosurveillance* *25*(10), 2000199.
- Li, L.-q., T. Huang, Y.-q. Wang, Z.-p. Wang, Y. Liang, T.-b. Huang, H.-y. Zhang, W. Sun, and Y. Wang (2020). Covid-19 patients’ clinical characteristics, discharge rate, and fatality rate of meta-analysis. *Journal of medical virology* *92*(6), 577–583.
- Liu, D., Y. Wang, J. Wang, J. Liu, Y. Yue, W. Liu, F. Zhang, and Z. Wang (2020). Characteristics and outcomes of a sample of patients with covid-19 identified through social media in wuhan, china: observational study. *Journal of medical Internet research* *22*(8), e20108.
- Liu, J., H. Nie, S. Li, X. Chen, H. Cao, J. Ren, I. Lee, and F. Xia (2021). Tracing the pace of covid-19 research: Topic modeling and evolution. *Big Data Research* *25*, 100236.
- Maneejuk, P., S. Thongkairat, and W. Srichaikul (2021). Time-varying co-movement analysis between covid-19 shocks and the energy markets using the markov switching dynamic copula approach. *Energy Reports*.
- Massicotte, P. and D. Eddelbuettel (2021). *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.4.8.
- Nagler, T., U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, and T. Erhardt (2021). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.4.2.
- Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

- O’Leary, D. E. and V. C. Storey (2020). A google–wikipedia–twitter model as a leading indicator of the numbers of coronavirus deaths. *Intelligent Systems in Accounting, Finance and Management* 27(3), 151–158.
- Peng, Z., R. Wang, L. Liu, and H. Wu (2020). Exploring urban spatial features of covid-19 transmission in wuhan based on social media data. *ISPRS International Journal of Geo-Information* 9(6), 402.
- Pinheiro, C. A. R., M. Galati, N. Summerville, and M. Lambrecht (2021). Using network analysis and machine learning to identify virus spread trends in covid-19. *Big Data Research*, 100242.
- Qin, L., Q. Sun, Y. Wang, K.-F. Wu, M. Chen, B.-C. Shia, and S.-Y. Wu (2020). Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index. *International journal of environmental research and public health* 17(7), 2365.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rahimi, I., A. H. Gandomi, P. G. Asteris, and F. Chen (2021). Analysis and prediction of covid-19 using sir, seiqr and machine learning models: Australia, italy and uk cases. *Information* 12(3), 109.
- Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3), 507–554.
- Rigby, R. A. and M. D. Stasinopoulos (1994). Robust fitting of an additive model for variance heterogeneity. In *Compstat*, pp. 263–268. Springer.
- Rigby, R. A., M. D. Stasinopoulos, G. Z. Heller, and F. De Bastiani (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press.
- Sifat, I., A. Ghafoor, and A. A. Mand (2021). The covid-19 pandemic and speculation in energy, precious metals, and agricultural futures. *Journal of Behavioral and Experimental Finance* 30, 100498.
- Silge, J. and D. Robinson (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Statistical Software* 1(3).

- Simard, C. and B. Rémillard (2015). Forecasting time series with multivariate copulas. *Dependence modeling* 3(1).
- Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8, 229–231.
- Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press.
- Wang, J., L. Wang, J. Xu, and Y. Peng (2021). Information needs mining of covid-19 in chinese online health communities. *Big Data Research* 24, 100193.
- Wynants, L., B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray, et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* 369.
- Zhu, Y., K.-W. Fu, K. A. Grépin, H. Liang, and I. C.-H. Fung (2020). Limited early warnings and public attention to coronavirus disease 2019 in china, january–february, 2020: a longitudinal cohort of randomly sampled weibo users. *Disaster medicine and public health preparedness* 14(5), e24–e27.