Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 84 No 6, 1324 doi: 10.2166/wst.2021.343

Data filtering methods for SARS-CoV-2 wastewater surveillance

Rezgar Arabzadeh[a], Daniel Martin Grünbacher[a], Heribert Insam[b], Norbert Kreuzinger[c], Rudolf Markt[b] and Wolfgang Rauch[a,*]

[a] Unit of Environmental Engineering, Department of Infrastructure, University of Innsbruck, Technikerstrasse 13, 6020 Innsbruck, Austria
[b] Department of Microbiology, University of Innsbruck, Innsbruck, Austria
[c] Institute for Water Quality and Resource Management, Technische Universität Wien, Vienna, Austria
*Corresponding author. E-mail: wolfgang.rauch@uibk.ac.at

WR, 0000-0002-6462-2832

ABSTRACT

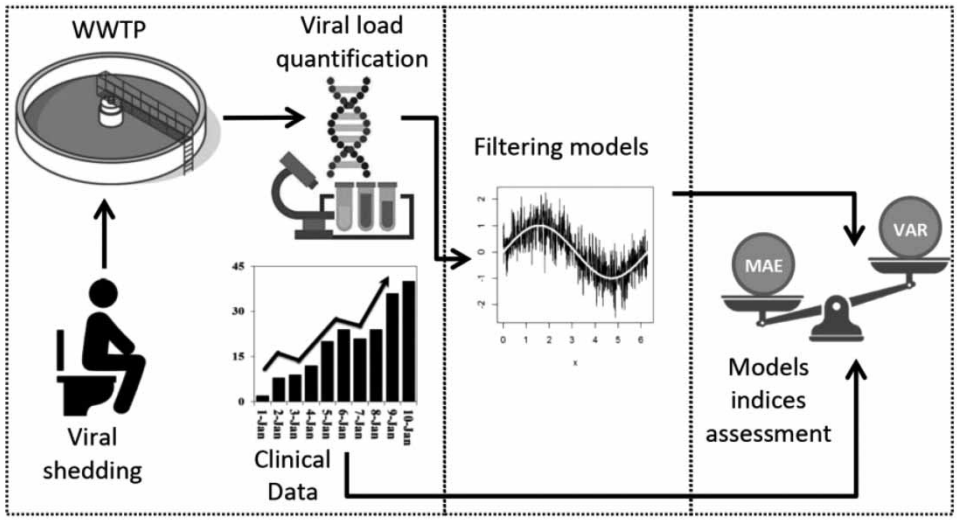
In the case of SARS-CoV-2 pandemic management, wastewater-based epidemiology aims to derive information on the infection dynamics by monitoring virus concentrations in the wastewater. However, due to the intrinsic random fluctuations of the viral signal in wastewater caused by several influencing factors that cannot be determined in detail (e.g. dilutions; number of people discharging; variations in virus excretion; water consumption per day; transport and fate processes in sewer system), the subsequent prevalence analysis may result in misleading conclusions. It is thus helpful to apply data filtering techniques to reduce the noise in the signal. In this paper we investigate 13 smoothing algorithms applied to the virus signals monitored in four wastewater treatment plants in Austria. The parameters of the algorithms have been defined by an optimization procedure aiming for performance metrics. The results are further investigated by means of a cluster analysis. While all algorithms are in principle applicable, SPLINE, Generalized Additive Model and Friedman's Super Smoother are recognized as superior methods in this context (with the latter two having a tendency to over-smoothing). A first analysis of the resulting datasets indicates the positive effect of filtering to the correlation of the viral signal to monitored incidence values.

Key words: data smoothing, pandemic management, SARS-CoV-2, signal filtering, virus monitoring, wastewater-based epidemiology

HIGHLIGHTS

- The random component in the timeline of SARS-CoV-2 virus concentration makes data filtering necessary.
- Thirteen common filtering techniques are investigated for their potential to smooth the virus signals.
- SPLINE, GAM and Friedman's Super Smoother are seen as superior algorithms for smoothing SARS-CoV-2 signals.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Management of the SARS-CoV-2 pandemic rests upon measures such as hygiene, isolation and vaccination but requires rigorous monitoring on the state and spread of the disease (Nicola *et al.* 2020). Next to individual qPCR and antigen testing, wastewater-based epidemiology (WBE) has been recognized as a valuable tool to estimate viral prevalence (Weidhaas *et al.* 2021). There is a rapidly increasing body of evidence about the methodology and its application (see e.g., Ahmed *et al.* 2020; Kitajima *et al.* 2020; Wu *et al.* 2020a, 2020b; Zhu *et al.* 2021). Several studies, e.g. He *et al.* (2020) or Wölfel *et al.* (2020) showed that infected persons in a catchment shed a certain amount of viral load per day into the sewer system, resulting in measurable viral titers at the sampling point – expressed as virus RNA concentrations [number of RNA copies/ml]. Such measurements are usually taken as composite samples at the treatment plant of the urban drainage system.

A key element in WBE is to derive information on the infection dynamics in the catchment by means of the monitored virus particle concentrations by quantifying their RNA genome. The signal serves as a proxy for prevalence, i.e., the total number of infected persons in the catchment (Ahmed *et al.* 2020). WBE is thus a valuable additional source of information next to individual testing strategies. Even more, time series prediction can be applied to the signal, thus serving as potential early-warning tool in pandemic management (e.g., Gonzalez *et al.* 2020; Hart & Halden 2020).

Relating to the basic concepts of WBE (Choi *et al.* 2018; Feng *et al.* 2018) it is typical not only to use the raw signal (RNA concentration as equivalent to virus particles) for further analysis but to apply normalization regarding (a) flow dynamics and (b) changes in population by use of biomarkers (Been *et al.* 2014). Still, it is due to the complexity of the whole process that the wastewater titer signals (both raw value and normalized) not only express the prevalence information but also contain huge variations. Reasons are manifold, but key factors are (a) the individual variances in viral shedding (both amount and time) of the infected persons, (b) effect of spatial distribution of the viral load in the catchment (i.e., the location of the main entry points), (c) stochastic influences to transport mechanics and virus degradation in the sewer system, (d) influence of rain runoff due to dilution and loss via CSO, and (e) variances that stem from both the sampling procedure and the laboratory methods. To conclude, the wastewater signal (even if normalized) contains not only the sought-after information on prevalence in the catchment but also a large noise contribution. The latter, however, makes the analysis and data interpretation difficult and data modeling a poorer fit to the actual status (Wand 2003; Samuelsson *et al.* 2017).

To differentiate the noise from the actual information in the signal, filtering techniques are frequently used in science and engineering (Huang *et al.* 2016). Simple filter techniques such as moving average are common, but recent studies (e.g., Stadler *et al.* (2020), Wu *et al.* (2020a, 2020b), Nemudryi *et al.* (2020), and Graham *et al.* (2020)) applied more advanced procedures such as locally weighted polynomial or spline methods to smooth the viral loads signals. It is clear that the choice of the filtering method – as data preprocessing step – is subject to the aim of the subsequent data use. Still, a thorough investigation to identify the optimal smoothing method applicable to filter SARS-Cov-2 time series is missing. In this study, 13 filtering

methods – from simple to advanced – have been applied to the viral load measured at four locations in Austria and are investigated towards three performance indicators, i.e., mean absolute error (MAE), variability (VAR), and Akaike Information Criterion (AIK).

In the remainder of the paper, we first outline the status of WBE in Austria, the selected four case studies and the titer datasets derived therefrom. The filtering methods are presented, however not elaborated in detail. To rationalize optimal performance, an optimization procedure is used for performance metric. The results are further investigated by means of a cluster analysis. Last, the explanatory power of the datasets with respect to infection dynamics is analyzed.

2. MATERIALS AND METHODS

2.1. Wastewater surveillance and datasets

Already early in the pandemic Austria has established the research project Coron-A to develop the scientific background of WBE as a Covid-19 surveillance tool (Coron-A 2021). Fundamental in the project is the surveillance of 23 wastewater treatment plants (WWTPs) in Austria by taking 24 h composite volume proportional samples (CVVT: constant volume; variable time) from the inlet of the WWTPs. The sampling frequency is bi-weekly or higher.

Sampling is done by cooled automatic samplers (various suppliers). Samples were cooled to 4 °C (Markt *et al.* 2021) and shipped to the laboratory in Styrofoam boxes with coolpacks guaranteeing continuous cooling during transport. In the laboratory, the 70 g sample was centrifuged for 30 min at 4,500 *g* (4508 R cooling centrifuge, Eppendorf, Hamburg, Germany) to remove particulate matter. The supernatant was then concentrated through polyethylene glycol (PEG) centrifugation at 12,000 *g* for 99 minutes. The pellet obtained was suspended in 800–1,000 µl lysis buffer (details see Markt *et al.* 2021) and transferred to a microreaction tube (Eppendorf). The RNA was purified using the Monarch™ total RNA Miniprep Kit (New England Biolabs, Ipswich, USA). After Nanodrop RNA quantification and appropriate dilution the SARS-CoV-2 nucleocapsid (N1) gene RNA copy numbers were determined on a RotorGene cycler (Qiagen, Hilden, Germany) using a plasmid standard containing the N gene of SARS-CoV-2 (2019-nCoV_N_Positive Control, IDT, Leuven, Belgium) (Markt *et al.* 2021). According to Pérez-Cataluña *et al.* (2021) the recovery can be estimated as approximately 50%.

According to national regulations, WWTPs apply a self-monitoring scheme and measure flow rate and temperature on a daily basis. Water quality parameters such as COD, N_{tot} and NH_4^+ are analyzed as well but the frequency depends on the design capacity of the investigated WWTP (varying between daily to weekly). Water quality parameters are likewise determined via the same 24 h composite samples as used for determination of the SARS-CoV-2 titer.

For our study, we selected four different sampling locations or urban drainage catchments respectively (see Table 1). For easier reference and respecting data protection acts in Austria, the catchments/cities are denoted as A–D in the following. The locations mainly vary in population and catchment area size as well as type of sewage system.

Case studies A and B represent prototypical Austrian cities (large to medium) with high population density and an urban environment. In both cases, the entity of the urban catchment is discharged to the WWTP. Case studies C and D, on the other hand, resemble smaller settlements and case study D is moreover a highly touristic place with predominately summer tourism. Figure 1 depicts the locations of the case studies within Austria's administrative regions. Meteorological data from 1971 to 2000 show a temperate climate for all sampling sites. However, the locations experience up to 40 days/year with a total daily precipitation of 10 mm or higher which leads to significant runoff and to a loss of virus particles in the sewage by combined sewer overflow.

Table 1 | Sampling locations – served population and climatic conditions

Sampling site		Connected residents	Avg. daily inflow 2020 (m3/d)	Avg. monthly temperature 1971–2000 (°C)[a]	Avg. total annual precipitation 1971–2000 (mm/a)[a]	Avg. number of days with total daily precipitation >10 mm (d/a)[a]
Urban	A	1900000	539450	11.4	548	14.9
	B	320681	83187	9.0	1184	40.0
Rural	C	41696	16344	8.9	1231	40.3
	D	23600	4899	7.9	889	29.1

[a]Zentralanstalt für Meteorologie und Geodynamik ZAMG (2002).

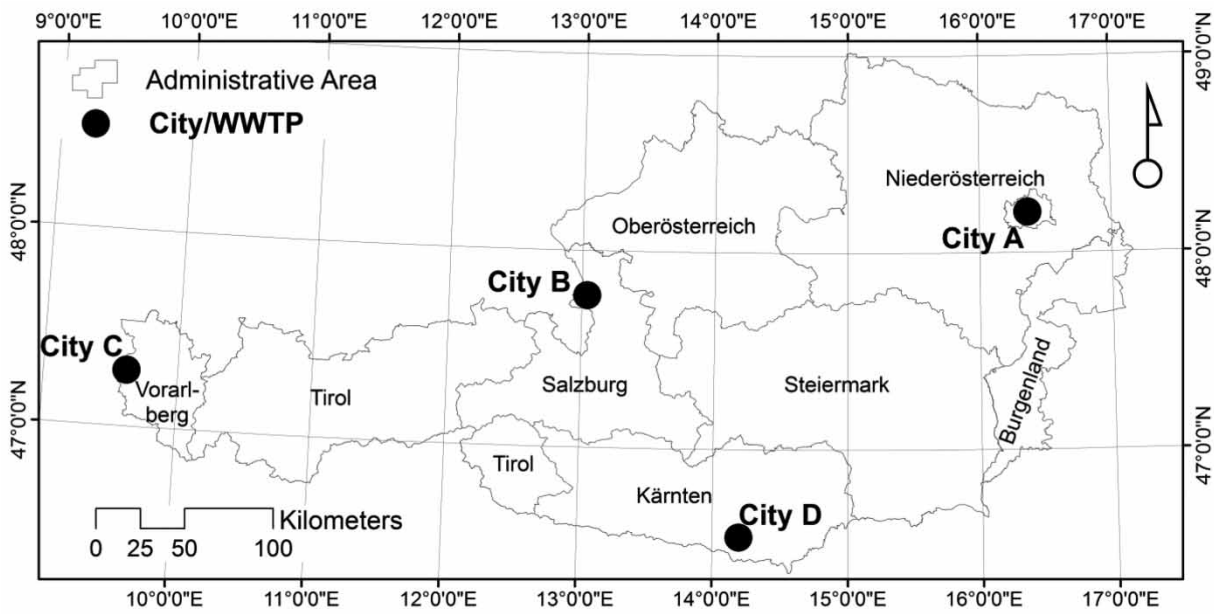


Figure 1 | Locations of case studies in Austria.

Wastewater surveillance in the four chosen case studies started in summer 2020 (in case of location A already in May 2020) and samples were taken weekly or more frequently. In this study we concern ourselves with the data until the end of 2020, which is a timeline of 8 months – see Figure 2 for details.

The timeline of the (raw) wastewater titer values follows the epidemic data derived from individual testing. For the latter we depict here active cases as identified by summing up infections versus recovered/deceased cases (Figure 2). The lockdown after the first pandemic wave in March 2020 was quite successful in reducing also the virus signal in wastewater. This is demonstrated by the low RNA concentration measured at city A in the early summer period. In principle, the signal remained at a low level for all sites during summer and early fall. One exception was case study D where a sharp increase in RNA concentrations was observed in August, potentially being related to the increase in summer tourism. Starting with October 2020 the beginning of the second Austrian pandemic wave was depicted also in the wastewater signal. Both the reported cases of infections and the viral RNA concentration in the four WWTPs peaked in mid-November. Thereafter, another lockdown has been imposed over the country that again declined both the infections and the RNA signal. Note that the temporal differences in the infection dynamics in the four case studies are likely to be due to the regional epidemic management.

2.2. Normalization

Wastewater-based epidemiology for pandemic management aims at deriving information on prevalence in the catchment. Prevalence is defined here as the fraction of the infected persons within the total population discharging into the sewer system. If – for the sake of simplicity – we assume that each infected person sheds a certain virus load per day, we can relate the measured virus concentration c_{virus} to the infection dynamics. We are thus less interested in the raw surveillance data but in the specific viral load instead and derive – for an arbitrary datapoint in the series:

$$L_{virus} = \frac{c_{virus} * Q}{P} \tag{1}$$

where L_{virus} is RNA copies/P/d; Q = flow volume in L/d; c_{virus} = virus concentration in the sample in RNA copies/L and P = number of persons in the watershed. While the consideration of flow in the timeline is evident from the measured inflow data at the WWTP (see section 2.1) the temporal variation of P in the catchment is to be estimated via a wastewater biomarker (Been *et al.* 2014) as:

$$P = \frac{c^{bm} * Q}{f_{bm}} \tag{2}$$

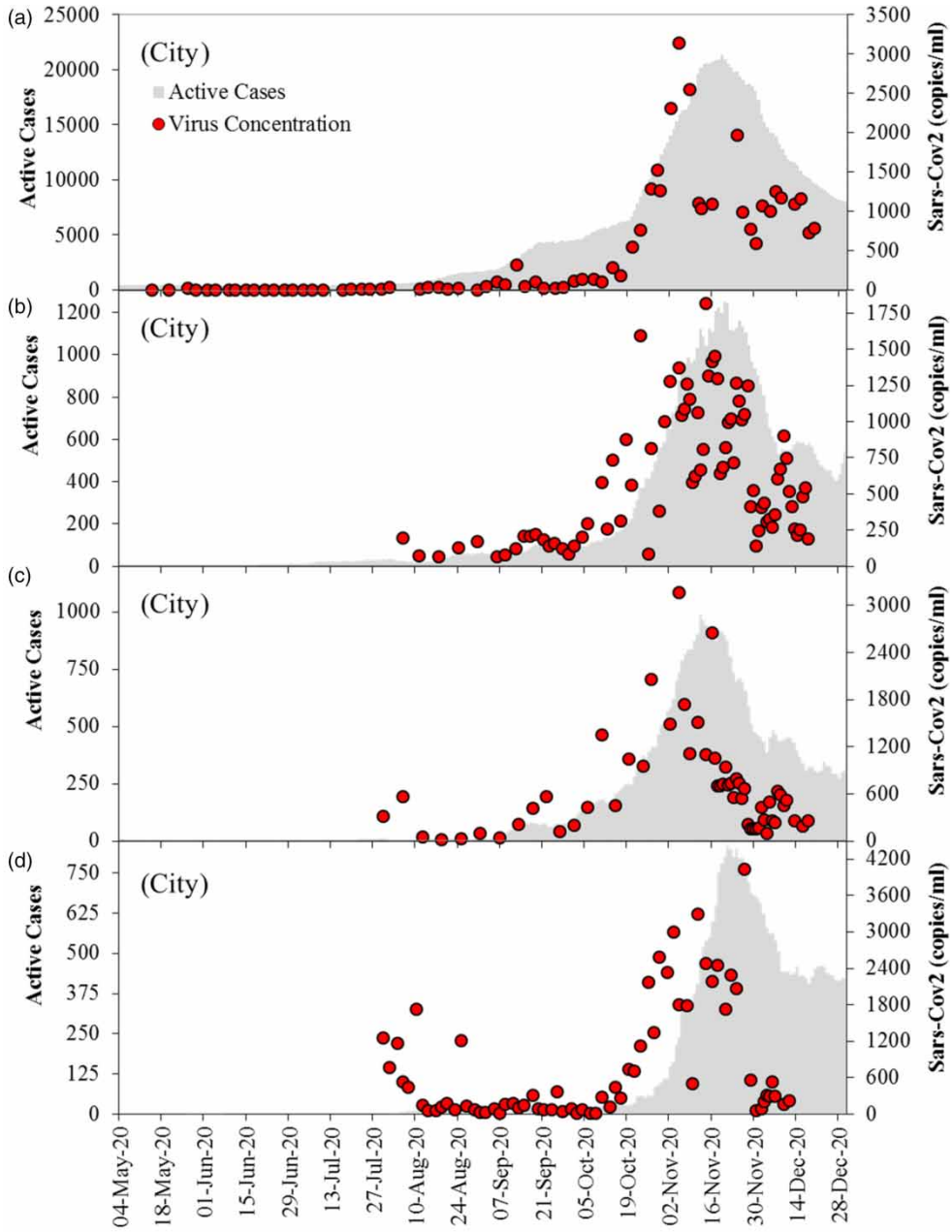


Figure 2 | Raw data timelines of SARS-CoV-2 titer values (RNA copies/ml) and epidemiological timelines (active cases as identified by PCR-tests) at the four sampling sites.

where c_{bm} is the concentration of biomarker in g/L and f_{bm} = specific biomarker load in g/P/d. The choice of an appropriate biomarker has been subject to numerous investigations (Choi *et al.* 2018). However, for this investigation we are less interested in actual values but can express the influence by normalization. As biomarker that is readily available at wastewater treatment plants due to regular surveillance, we apply the standard water quality parameter NH_4-N. The specific load f_{bm} is here derived from the measured 50-percentile value in the period of the first lockdown in Austria as load fluctuations are minimal therein. Despite NH_4-N being potentially influenced by industry contribution, the parameter is applicable in this context (Been *et al.* 2014; Rauch *et al.* 2021)

2.3. Filtering techniques

The timeline of surveillance raw data (and normalized data as well) includes not only the sought-after information regarding prevalence in the catchment but contains a significant noise contribution, that is due to stochastic effects in the whole

process. In this study we apply and compare 13 filter/smoothing techniques with the aim to de-noise the time series of RNA-concentration in WBE. Table 3 summarizes the methods applied herein and gives the key reference(s) for each.

Typically, the filtering methods can be discriminated by the parameters needed for its use, ranging between 0 and more than 3. Parameter-less data models (here FFT, TUK, and KAF) are mathematically complex and (usually) computationally more difficult to implement, the benefit being obvious as calibration is omitted for these techniques. Data models typically have one to three adjustable parameters. The most common used parametric method in the engineering community is (centralized) SMA which is both simple to implement and contains one parameter only. The detriment is the lower robustness as compared to other – more complex – methods (Williams *et al.* 1998; Raudys *et al.* 2013). The GAM model, which is a more recent development, contains a high number of parameters and is a combination of additive and generalized linear models (Wood *et al.* 2016). Since the parameters can be estimated, the GAM method could also be applied as parameter-less (but not done herein). The numerical details of methods implemented in this study are omitted as this information is given exhaustively in the literature (see reference column in Table 2). The methods are implemented in R and have been tested prior to application for reference datasets.

Some of the implemented models are not only used for smoothing but also provide means for signal forecasting, e.g., GAM, ADP, ARI, POL, SUP, KER, and SPL. Since this study focuses on signal filtering, the capability and accuracy of the techniques in terms of prediction are not taken into account.

2.4. Workflow

The general workflow for the investigation is depicted in Figure 3. The overall data analysis is divided into two main steps a) model fitting and b) clustering. In the first step, multiple smoothing algorithms are applied to the SARS-CoV-2 titer values (raw and normalized to NH_4) to de-noise the measured signals in the wastewater system. To quantify their performance a cross-validation approach (Stone 1978) is implemented to estimate a precise error value associated with each model configuration. First, a given filter is fitted to the SARS-CoV-2 time series ($x_{[1,2,\ldots,T]}$) under the absence of any arbitrary entity of x and the resulting model fitness and metrics are calculated. This procedure is repeated by subsequent excluding each single remaining entity of the series, until a T-by-T fitness matrix of $\hat{x}$ is computed. Thus, for any observed measurement, there will be T fitted values available, i.e., vector $\hat{x}_{t*}$. From $\hat{x}_{t*}$ we compute the model prediction including uncertainty bounds by using the empirical cumulative distribution function of $\hat{x}_{t*}$.

As most models contain adjustable parameters, calibration is an essential step of the workflow. We apply mathematical optimization (as needed either in discrete or continuous mode) to guarantee the best model structure/parameter selection. As stated frequently in the literature the genetic algorithm (GA) is well adapted to solve both real-valued or integer programming even for complex and ill-posed problems (e.g., Panchal & Panchal 2015).

The second step involves clustering of the methods according to their performance (error and consistency). As some of the methods are functionally similar, the center of the clusters is seen as a representative solution. For clustering the K-Medoid algorithm (Park & Jun 2009) is applied.

2.5. Calibration of parametric data models

Filtering methods with parameter(s) require calibration to estimate the best configuration. Global calibration algorithms based on mathematical optimizers are manifold in science and engineering (Schutte *et al.* 2004; Price *et al.* 2006; Kaur *et al.* 2020). Since GA has been applied successfully to different optimization problems (Mehr *et al.* 2018; Ghoddusi *et al.* 2019) and is suitable to solve either permutation (integer) programming or real-valued optimization as required herein, GA was selected as optimization procedure (see Mirjalili (2019) about GA details and operators). Table 4 summarizes the parameters/configurations used for GA to calibrate the parametric filtering methods.

Table 2 | Specific NH_4-N load per person per day

	City A	City B	City C	City D
2.5% percentile	9.77	5.84	8.02	5.94
50.0% percentile	10.71	6.49	8.99	6.80
97.5% percentile	12.17	7.13	9.73	9.32

Calculated percentiles from daily measurements during the first lockdown period in Austria (April to mid-May 2020).

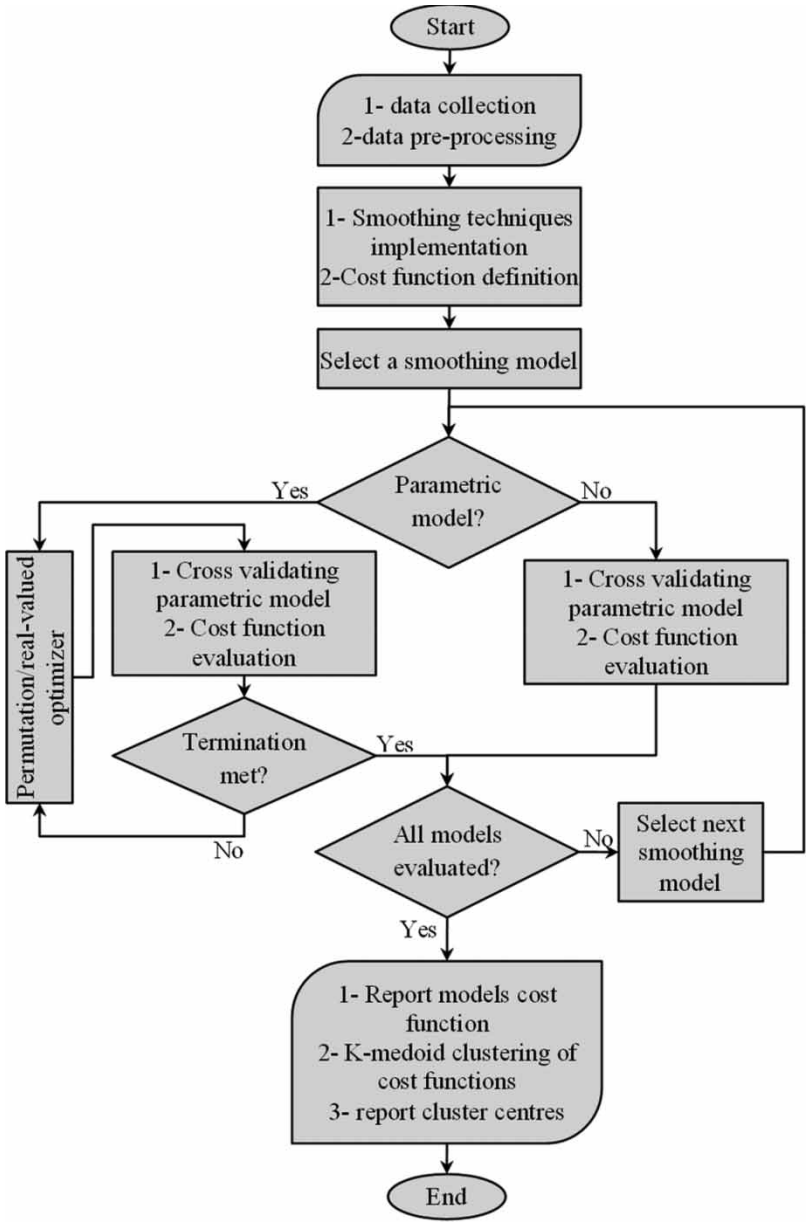


Figure 3 | Flowchart for analysis of smoothing methods.

2.6. Performance indices

To measure the robustness of the filtering methods against the absence of signal entities, a multi- criteria indexing approach is performed to address the validity of methods for smoothing SARS-CoV-2 time series data. To this end, a cost function comprised of mean absolute error (MAE), variability (VAR), and Akaike information criterion (AIC) (Sakamoto *et al.* 1986; Anderson & Burnham 2004) was computed for every model. With x_t as the original signal and $\hat{x}$ as a column-wise square t-by-t matrix of the filtered series (where each column represents filtered values of x_t under the absence of the t^{th} signal value) the performance indicators MAE, VAR, and AIC are computed as:

$$MAE = \frac{\sum_t |diag\{\hat{x}\}_t - x_t|}{T} \tag{3}$$

$$VAR = \sum_t \frac{\sum_t (\hat{x}_{t*} - \overline{\hat{x}_{t*}})}{T-1} \tag{4}$$

Table 3 | Applied smoothing algorithms

Adaptive Degree Polynomial Filter (ADP)
Auto Regressive Model (ARI)
Fast Fourier Transform Filtering (FFT)
Friedman's Super Smoother (SUP)
Generalized Additive Model (GAM)
Kalman Filtering (KAF)
Kernel Smoother (KER)
Locally-Weighted Polynomial (POL)
Robust Running Medians (RRM)
Savitzky-Golay Filters (SGF)
Simple Moving Average (SMA)
Spline (SPL)
Tukey Smoother (TUK)

Method	Reference	Sample
TUK	Mallows (1979)	Fiskeaux & Ling (1982)
KAF	Tusell (2011)	Pan *et al.* (2016)
FFT	Cochran *et al.* (1967)	Yang *et al.* (2004)
SPL[a,e]	Reinsch (1967)	Eubank (1988)
KER[a,e]	Härdle & Vieu (1992)	Speckman (1988))
SMA[a]	Hyndman (2011)	He et al. (2020)
RRM[a]	Friedman & Stuetzle (1982)	Polasek (1984)
SUP[a,e]	Friedman (1984)	Friedman & Silverman (1989)
POL[a,e]	Atkeson *et al.* (1997)	Rajagopalan & Lall (1998)
SGF[b]	Press & Teukolsky (1990)	Bromba & Ziegler (1981)
ARI[b,e]	Akaike (1969)	Lohani *et al.* (2012)
ADP[c,e]	Barak (1995)	Jakubowska & Kubiak (2004)
GAM[d,e]	Hastie (2017)	Murphy *et al.* (2019)

[a]Single parameter.
[b]Double parameter.
[c]Triple parameter.
[d]Above triple parameter model.
[e]Models with the ability of forecasting.

Table 4 | GA parameters used to calibrate filtering models

Parameter/configuration	Value/method
Population size	100
Iteration	1000
Mutation rate	0.1
Crossover rate	0.8
Elitism	0.05
Selection	roulette wheel
Mutation method	Random
Crossover	Two points

$$AIC = T.\log\left(\frac{\sum_t (diag\{\hat{x}\}_t - x_t)^2}{T}\right) + 2k \tag{5}$$

where $\hat{x}_{t*}$ is the t^{th} row of $\hat{x}$ matrix, $\overline{\hat{x}_{t*}}$ is average of $\hat{x}_{t*}$, k is the number of model parameters and $1 < t < T$.

3. RESULTS AND DISCUSSION

Applying the methods described in section 2, the suitability of smoothing methods is tested for the viral load signals of the four case studies. Note that we first apply the whole procedure as described in 2.4 to the raw signal and – in a second step – repeat the procedure for the NH_4 normalized signal.

3.1. Raw signal

Figure 4 shows the results of 13 implemented filtering techniques for the indicators MAE and VAR. According to the results, the MAE varies significantly from station to the station, while the indicator VAR varies similarly across stations. Generally, we can see a strong influence of the catchment size (city A-D) in the results. For the dataset City A (large city), both MAE and VAR are smaller for all filtering techniques as compared to the values obtained for the dataset City D (small community). City B and City C corroborate the trend, that MAR and VAR are increasing the smaller the catchment size. In Figure 4, the methods separated with a solid box indicate the center of the K-Medoid clustering for each station.

For the K-Medoid algorithm, performance indices have been partitioned into 3 clusters namely as best, middle and worst. Next, the median of the best clusters has been defined as optimal method. As a result, for the raw signal investigation, we found the SPLINE method to exhibit best results in both MAE and VAR for the datasets A, B and C. Only in the dataset for the smallest (and touristic) catchment (City D) it is the Locally-weighted polynomial method that behaves best. However, note that also for dataset D Spline is clustered among the best.

Table 5 shows the results of clustering for the filtering methods. Accordingly, the majority of the methods are placed in cluster 1, that indicates the best clustered methods. It is worth to mention that Kalman filtering (KAL) performs worst for smoothing of the viral signals in all WWTPs. Conversely, Friedman's Super Smoother (SUP), Spline (SPL) and Savitzky-Golay Filters (SGF) outweigh the other filtering methods in most of the cases. According to Table 5, SPL and SUP as parametric and nonparametric methods, respectively, are the only ones among the best filtering techniques in all four WWTPs. Two-parameter methods, ARI and SGF, are mostly clustered in the best cluster, while nonparametric methods such as TUK and FFT have been listed in both the best and the second-best clusters with the same membership frequencies.

Figure 5 indicates the raw signal of viral RNA concentration measured in the four wastewater treatment plants and the smoothed signal as computed by the optimal method (center method of the best cluster). Also the result of the cross-validation is depicted as well as the 95% confidence interval to assess the uncertainty of the filtering. Note that the computed uncertainty bounds are higher for City D (small population) and lower for City A (large population). Uncertainty increases

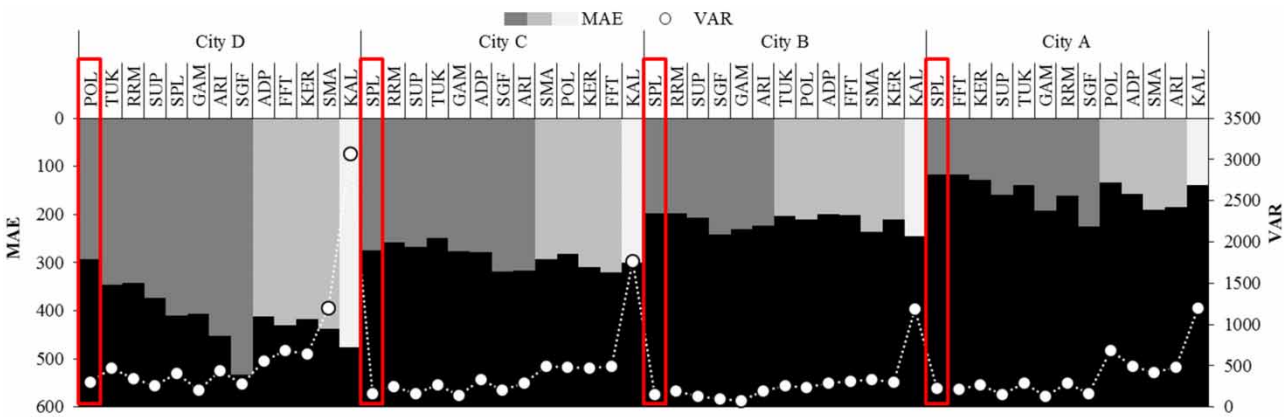


Figure 4 | Performance of filtering methods for raw signals (black bars: mean absolute error (MAE) and white dots: variance (VAR)). Optimal methods are placed left for each case study.

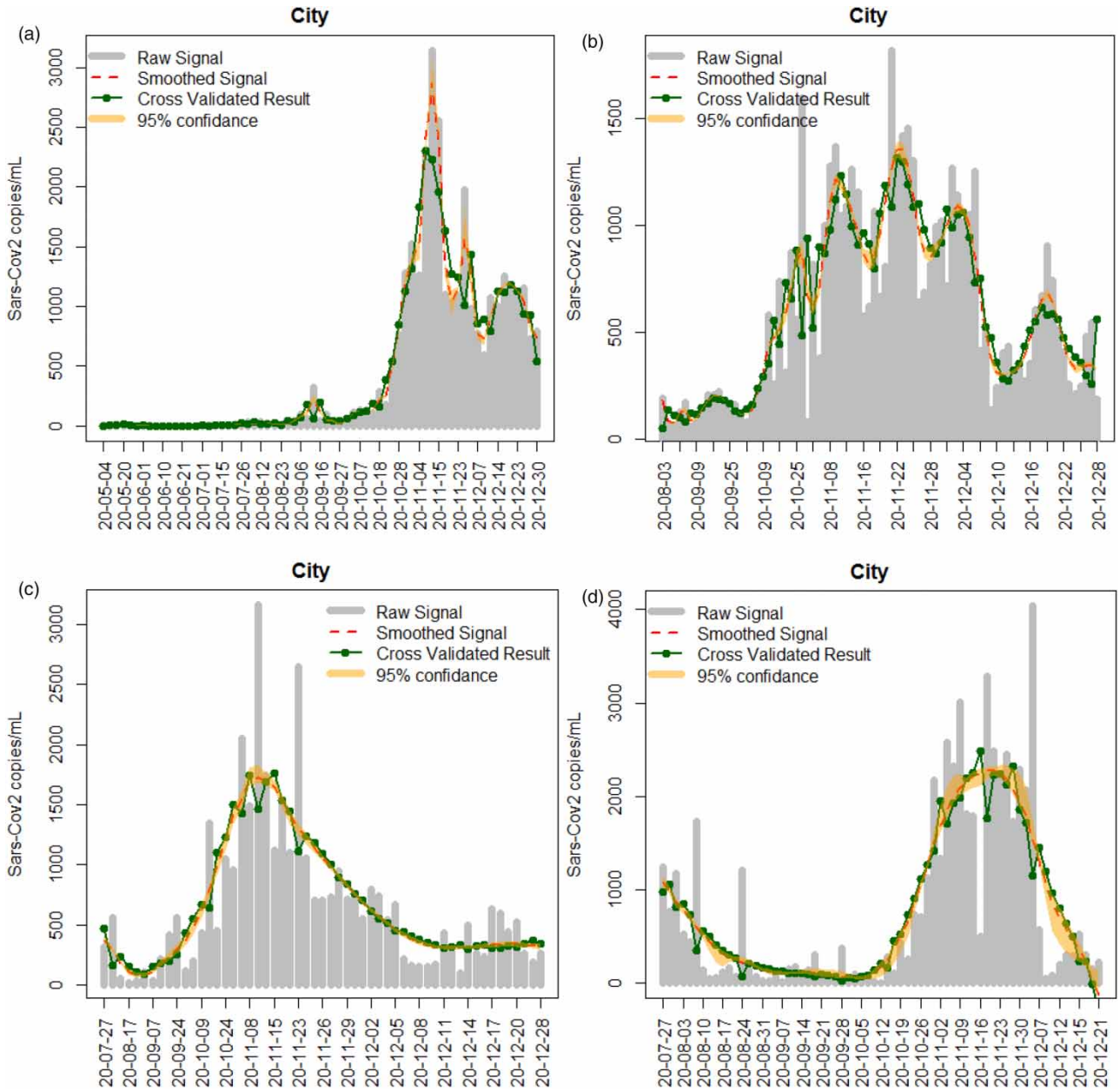


Figure 5 | Application of the selected smoothing techniques to the raw data signal (superimposed with the result of cross-validation and the 95% confidence interval).

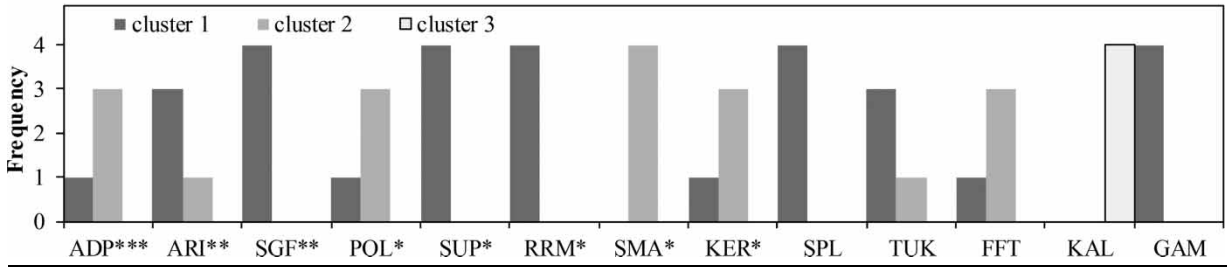
where there is a peak (positive or negative) in the original data series or where a significant difference is seen between the original signal and the smoothed one.

3.2. Normalized signal

Applying the procedure outlined above to the SARS-CoV-2 signal normalized to NH_4, the general picture does not change drastically. For clustering the methods according to their performance, the K-Medoid algorithm was applied as well – see Table 6. As a result, POL, SPL, SUP, and GAM are identified as optimal smoothing methods. Although the optimal results (center of the best cluster) are different for the two signals (raw versus normalized), the methods SPL, GAM, SUP, and SGF are found as suitable/preferable for both signals. It is also notable that KAL again is consistently performing worst among the deployed methods.

Table 5 | Clustering of performance indices for raw signal investigation

City D				City C				City B				City A			
Method	VAR	MAE	AIC	Method	VAR	MAE	AIC	Method	VAR	MAE	AIC	Method	VAR	MAE	AIC
POL	**305.9**	**292.9**	**846.6**	SPL	**159.0**	**275.3**	**662.7**	RRM	196.5	198.1	887.1	FFT	220.4	116.2	723.0
TUK	475.4	346.5	860.6	RRM	247.6	258.5	663.8	SPL	**154.6**	**198.3**	**889.6**	SPL	**229.8**	**117.3**	**729.9**
RRM	349.4	343.0	860.9	SUP	167.7	267.0	664.3	SUP	135.3	207.0	892.8	KER	265.5	128.9	743.9
SUP	261.5	373.8	861.6	TUK	274.5	249.7	664.4	SGF	100.7	241.5	909.2	SUP	146.9	158.9	778.0
SPL	410.8	410.3	866.4	GAM	146.2	276.7	671.3	GAM	74.8	231.2	909.8	TUK	285.8	139.8	784.1
GAM	202.8	407.1	873.2	ADP	329.1	279.0	671.4	ARI	192.6	223.4	915.8	GAM	132.9	193.1	791.9
ARI	440.4	451.9	883.6	SGF	201.8	317.7	679.2	TUK	253.7	202.6	890.9	RRM	293.4	161.3	795.3
SGF	277.2	533.2	890.6	ARI	295.3	316.4	684.1	POL	232.4	210.1	892.0	SGF	158.6	225.9	814.1
ADP	560.1	412.6	871.4	SMA	494.3	293.0	668.3	ADP	295.8	199.5	895.2	POL	688.5	133.0	747.3
FFT	687.6	429.5	875.4	POL	487.5	281.0	675.2	FFT	309.2	201.4	905.9	ADP	492.6	157.7	783.3
KER	644.7	417.8	875.8	KER	470.6	309.0	680.3	SMA	333.6	236.6	909.4	SMA	415.8	189.4	789.8
SMA	1195.0	436.7	879.6	FFT	489.9	319.5	681.4	KER	306.1	210.2	912.9	ARI	478.1	185.2	803.4
KAL	3063.9	476.4	885.7	KAL	1762.8	300.6	672.8	KAL	1186.3	245.0	921.0	KAL	1198.8	138.9	746.4



Similar to above, all normalized time series are depicted by using the optimal signal filtering method based on the clustering (Figure 6). Despite the difference in scale, the shape of the smoothed data is quite similar for both normalized and raw signals. A difference is the higher variation of the cross-validation results – indicating a higher sensitivity of the selected filtering techniques when applied to the normalized signal as compared to the raw data.

3.3. Virological analysis

Additional to the analysis of the smoothing methods, the dataset also allows to reflect on the relation of the wastewater measurements to the infection dynamics. As a ballpark approximation we plot the ratio of the viral load per capita to the 7-day incidence value (expressed as sum of new infections over 7 days per 100,000 persons) by fitting linear models to each case study for both filtered and raw viral loads (Figure 7). While this relation is a severe simplification disregarding the different statistical properties of the data, it still allows to reflect on the benefit of data filtering.

Using the marginal probability densities, we derive for the 50-percentile value of the viral load (appr. 100×10^6 copies/cap) incidence values ranging between 100 and 600 (7 day infections/100000 persons). The interesting feature is the influence of scale: From the linear models we see both higher intercepts and lower slopes for small catchments and vice versa. The first observation indicates that small catchments have a certain threshold of viral load before a clear relation with infection dynamics is seen. The second observation (i.e., the slope of the model) points to the fact that infection clusters are (statistically) more significant in a smaller population than in a large one. While still speculative, community size could be an influential factor for WBE in the case of SARS-CoV-2.

Further, the analysis allows demonstration of the benefit of data filtering. A first empirical indication is the narrowing in the model uncertainty bounds when comparing unfiltered data (Figure 7(a)) against filtered one (Figure 7(b)). A quantitative argument towards filtering is given by the increase of Pearson correlation coefficient between the datasets – once filtering is applied (Table 7). For all case studies a significant increase of the correlation coefficient is apparent, indicating the role of time series filtering in the enhancement of model performance quality.

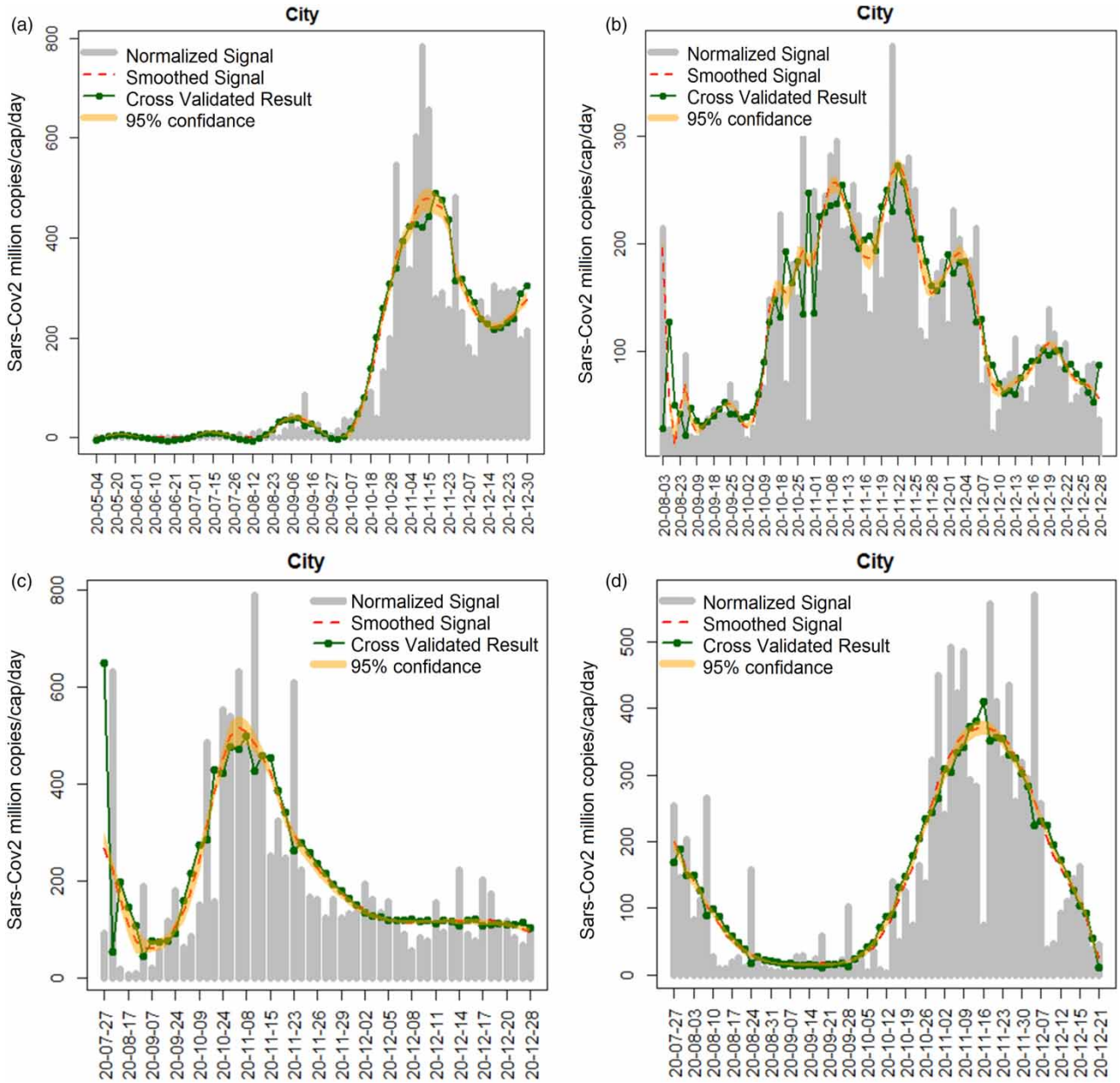


Figure 6 | Application of the selected smoothing techniques to the normalized data signal (expressed as RNA 10^6 copies/cap) superimposed with the result of cross-validation and the 95% confidence interval.

4. CONCLUSION

For management of pandemics such SARS-CoV-2 and interpretation of data obtained by means of WBE, filtering of the wastewater titer signal is an important pre-processing step. Modeling the infection dynamics or developing predictive tools therefore may induce misleading results when based on noisy information. This is especially important when models are not robust enough against extreme/oscillative inputs. This study focused on the application of 13 well-established signal filtering techniques for smoothing SARS-CoV-2 datasets in four wastewater treatment plants across Austria. Based on the finding in this study the following conclusions are made:

- Spline, GAM, and Friedman's Super Smoother are recognized as superior methods in this context. In three wastewater treatment plants Spline was found to be a robust approach to cope with missing data and uncertainties.

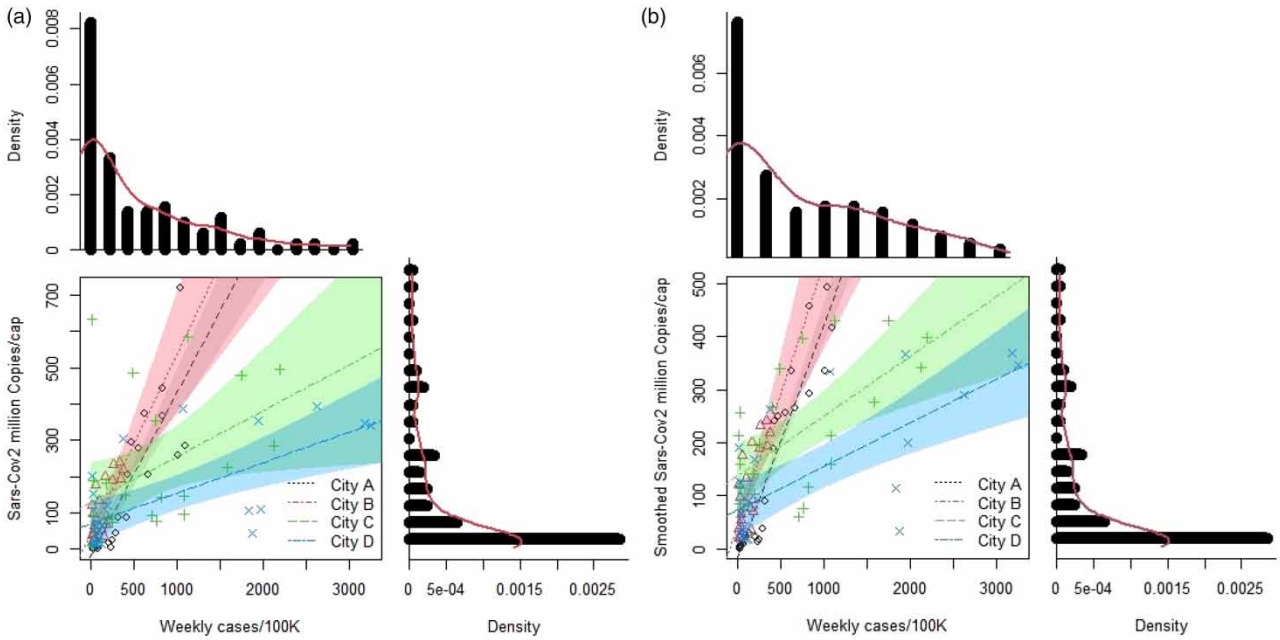


Figure 7 | Linear models expressing the ratio of normalized viral loads against 7-day incidence per 100,000 persons. (a) Non-filtered load against incidences. (b) Filtered load against incidences.

Table 6 | Clustering of performance indices for the normalized signal investigation

City D				City C				City B				City A			
Method	VAR	MAE	AIC	Method	VAR	MAE	AIC	Method	VAR	MAE	AIC	Method	VAR	MAE	AIC
GAM	**60.4**	**123.7**	**711.8**	SUP	**56.1**	**89.5**	**544.7**	SPL	**14.0**	**45.7**	**670.9**	POL	**69.3**	**31.8**	**550.1**
SGF	78.6	127.1	707.1	SGF	64.0	96.1	548.6	GAM	14.0	45.6	680.9	KER	56.3	33.4	558.2
SUP	102.5	121.0	703.3	ADP	81.5	83.1	542.4	SUP	24.2	46.4	677.1	FFT	62.6	34.2	554.4
ARI	133.2	135.2	719.7	TUK	84.8	78.3	543.4	ARI	26.9	47.7	687.8	SPL	63.8	35.4	563.2
SPL	135.4	123.3	704.6	GAM	88.4	98.8	564.9	SGF	29.1	47.2	680.3	KAL	261.6	40.6	581.4
RRM	142.4	108.3	699.3	RRM	88.9	79.1	544.5	RRM	40.7	44.6	677.0	TUK	85.0	41.8	611.1
SMA	143.2	111.0	694.8	SMA	125.3	97.6	560.2	ADP	52.1	47.2	685.5	SMA	80.0	43.0	593.9
ADP	161.5	125.4	706.8	ARI	144.6	120.4	576.2	TUK	55.4	46.4	681.8	RRM	105.0	44.5	613.0
TUK	163.1	112.7	704.7	KER	164.6	107.0	566.7	SMA	63.9	49.5	683.9	SUP	37.0	44.7	596.7
KER	204.5	127.6	719.6	SPL	196.5	109.7	568.5	KER	83.4	53.8	707.4	SGF	42.6	47.0	603.2
FFT	205.6	124.7	713.7	FFT	218.2	117.5	592.4	FFT	83.6	50.8	699.3	ADP	91.7	50.0	618.1
POL	247.7	117.5	713.3	POL	218.2	117.5	594.4	POL	83.6	50.8	701.3	GAM	34.2	51.3	611.0
KAL	723.0	129.1	703.1	KAL	481.3	106.1	559.7	KAL	203.0	49.1	687.0	ARI	108.3	51.8	630.4

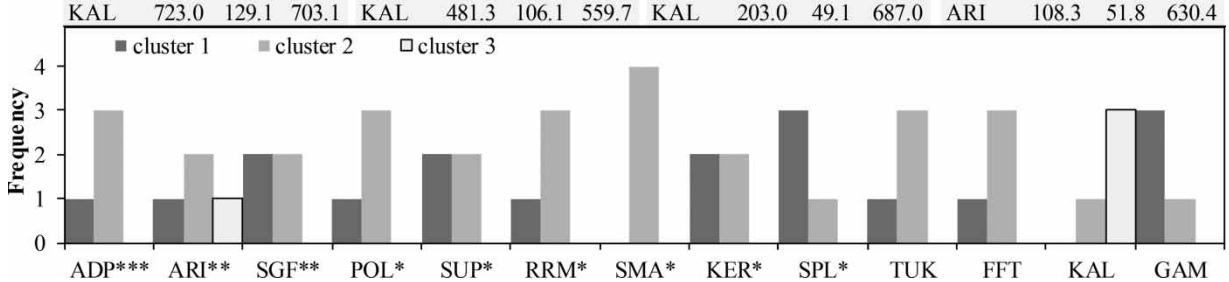


Table 7 | Squared Pearson correlation coefficients between incidences and raw and filtered viral loads

Sites	Raw	Filtered
City A	0.788[b]	0.920[b]
City B	0.601[b]	0.684[b]
City C	0.200[a]	0.394[b]
City D	0.453[b]	0.515[b]

[a]95% confidance.
[b]99% confidance.

- Although GAM is a robust smoother against extremes and outliers, it requires a high number of parameters to be tuned and has a tendency to over-smooth signals. The latter also applies to the Friedman's Super Smoother technique.
- For the case of nonparametric methods, TUK and FFT performed generally well and are suitable algorithms. However, nonparametric methods are sensible for missing values and are thus only recommended for times series with a small number of missing signals.
- Despite acceptable error values for methods such as KAL, SMA, and POL, they are not suitable in this context as generally overfitting.
- A first analysis of the dataset indicates that community size has an influence on WBE for SARS-CoV-2. For smaller catchments both a threshold of viral load is apparent before any relation with infection dynamics is visible and also a higher sensitivity towards infection clusters.
- For the application of linear regression models for incidence prediction, filtering results in a consistently improved Pearson correlation coefficient, i.e., model performance.

Apart from the applicability of GAM, Spline and POL (known as LOWESS), in filtering of the wastewater SARS-CoV-2 signals, the conclusions made herein are data-driven based and applying them to similar case cases must be made with discretion.

ACKNOWLEDGEMENTS

This study was funded by the Austrian Federal Ministry of Education, Science and Research, the Austrian Federal Ministry of Agriculture, Regions and Tourism, the federal states Burgenland, Carinthia, Lower Austria, Salzburg, Styria, Tirol, Upper Austria and Vorarlberg and the Austrian Association of Cities and Towns. We would like to thank the staff of the treatment plants involved for their support and the use of the wastewater titer measurement data. The support of Günther Weichlinger and Christoph Scheffknecht are also appreciated.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., Choi, P. M., Kitajima, M., Simpson, S. L., Li, J., Tscharke, B., Verhagen, R., Smith, W. J. M., Zaugg, J., Dierens, L., Hugenholtz, P., Thomas, K. V. & Mueller, J. M. 2020 First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of the Total Environment* **728**, 138764.

Akaike, H. 1969 Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21** (1), 243–247.

Anderson, D. & Burnham, K. 2004 *Model Selection and Multi-Model Inference. Second.* Springer-Verlag, NY, Vol. 63(2020), p. 10.

Atkeson, C. G., Moore, A. W. & Schaal, S. 1997 *'Locally Weighted Learning.' Lazy Learning*. Springer, Dordrecht, pp. 11–73.

Barak, P. 1995 Smoothing and differentiation by an adaptive-degree polynomial filter. *Analytical Chemistry* **67** (17), 2758–2762.

Been, F., Rossi, L., Ort, C., Rudaz, S., Delémont, O. & Esseiva, P. 2014 Population normalization with ammonium in wastewater-based epidemiology: application to illicit drug monitoring. *Environmental Science & Technology* **48** (14), 8162–8169.

Bromba, M. U. & Ziegler, H. 1981 Application hints for savitzky-Golay digital smoothing filters. *Analytical Chemistry* **53** (11), 1583–1586.

Choi, P. M., Tscharke, B. J., Donner, E., O'Brien, J. W., Grant, S. C., Kaserzon, S. L., Mackie, R., O'Malley, E., Crosbie, N. D., Thomas, K. V. & Mueller, J. F. 2018 Wastewater-based epidemiology biomarkers: past, present and future. *TrAC Trends in Analytical Chemistry* **105**, 453–469.

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., Maling, G. C., Nelson, D. E., Rader, C. M. & Welch, P. D. 1967 What is the fast Fourier transform? *Proceedings of the IEEE* **55** (10), 1664–1674.

Coron-A – Detection of corona virus in wastewater. Available from: https://www.coron-a.at/ (accessed 17 January 2021).

Eubank, R. L. 1988 *Spline Smoothing and Nonparametric Regression*, Vol. 90. Marcel Dekker, New York.

Feng, L., Zhang, W. & Li, X. 2018 Monitoring of regional drug abuse through wastewater-based epidemiology – a critical review. *Science China Earth Sciences* **61** (3), 239–255.

Fiskeaux, C. D. & Ling, R. F. 1982 Tukey smoothers as preprocessors for positive ar(1) parameter estimation in the presence of additive contamination. *Journal of Statistical Computation and Simulation* **15** (4), 315–331.

Friedman, J. H. 1984 *A Variable Span Scatterplot Smoother. Laboratory for Computational Statistics, Stanford University Technical Report No. 5*.

Friedman, J. H. & Silverman, B. W. 1989 Flexible parsimonious smoothing and additive modeling. *Technometrics* **31** (1), 3–21.

Friedman, J. H. & Stuetzle, W. 1982 *Smoothing of Scatterplots.* Stanford University California Project Orion.

Ghoddusi, H., Creamer, G. G. & Rafizadeh, N. 2019 Machine learning in energy economics and finance: a review. *Energy Economics* **81**, 709–727.

Gonzalez, R., Curtis, K., Bivins, A., Bibby, K., Weir, M. H., Yetka, K., Thompson, H., Keeling, D., Mitchell, J. & Gonzalez, D. 2020 COVID-19 surveillance in southeastern Virginia using wastewater-based epidemiology. *Water Research* **186**, 116296.

Graham, K. E., Loeb, S. K., Wolfe, M. K., Catoe, D., Sinnott-Armstrong, N., Kim, S., Yamahara, K. M., Sassoubre, L. M., Mendoza Grijalva, L. M., Roldan-Hernandez, L. & Langenfeld, K. 2020 SARS-CoV-2 RNA in wastewater settled solids is associated with COVID-19 cases in a large urban sewershed. *Environmental Science & Technology* **55** (1), 488–498.

Härdle, W. & Vieu, P. 1992 Kernel regression smoothing of time series. Journal of Time Series Analysis 13 (3), 209–232.

Hart, O. E. & Halden, R. U. 2020 Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: feasibility, economy, opportunities and challenges. *Science of the Total Environment* **730**, 138875.

Hastie, T. J. 2017 *Generalized Additive Models*. Boca Raton, Routledge, pp. 249–307.

He, Y., Wang, X., He, H., Zhai, J. & Wang, B. 2020 Moving Average Based Index for Judging the Peak of the COVID-19 Epidemic. *Int. J. Environ. Res. Public Health* **17**, 5288.

Huang, Y., Zhang, Y., Li, N. & Chambers, J. 2016 Robust student's t based nonlinear filter and smoother. *IEEE Transactions on Aerospace and Electronic Systems* **52** (5), 2586–2596.

Hyndman, R. J. 2011 Moving Averages. Contribution to the International Encyclopedia of Statistical Science, ed. Miodrag Lovric, Springer. pp. 866–869m.

Jakubowska, M. & Kubiak, W. W. 2004 Adaptive-degree polynomial filter for voltammetric signals. *Analytica Chimica Acta* **512** (2), 241–250.

Kaur, S., Awasthi, L. K., Sangal, A. L. & Dhiman, G. 2020 Tunicate swarm algorithm: a new bio-inspired based metaheuristic paradigm for global optimization. *Engineering Applications of Artificial Intelligence* **90**, 103541.

Kitajima, M., Ahmed, W., Bibby, K., Carducci, A., Gerba, C. P., Hamilton, K. A., Haramoto, E. & Rose, J. B. 2020 SARS-CoV-2 in wastewater: state of the knowledge and research needs. *Science of the Total Environment* **739**, 139076.

Lohani, A. K., Kumar, R. & Singh, R. D. 2012 Hydrological time series modeling: a comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. *Journal of Hydrology* **442**, 23–35.

Mallows, C. L. 1979 *Some Theoretical Results on Tukey's 3R Smoother.' Smoothing Techniques for Curve Estimation*. Springer, Berlin, Heidelberg, pp. 77–90.

Markt, R., Mayr, M., Peer, E., Wagner, A. O., Lackner, N. & Insam, H. 2021 Detection and stability of SARS-CoV-2 fragments in wastewater: Impact of storage temperature. medRχiv (Preprint). https://doi.org/10.1101/2021.02.22.21250768.

Mehr, A. D., Nourani, V., Kahya, E., Hrnjica, B., Sattar, A. M. & Yaseen, Z. M. 2018 Genetic programming in water resources engineering: a state-of-the-art review. *Journal of Hydrology* **566**, 643–667.

Mirjalili, S. 2019 Genetic algorithm. In: *Evolutionary Algorithms and Neural Networks*. Springer, Chambery, pp. 43–55.

Murphy, R. R., Perry, E., Harcum, J. & Keisman, J. 2019 A generalized additive model approach to evaluating water quality: Chesapeake Bay case study. *Environmental Modelling & Software* **118**, 1–13.

Nemudryi, A., Nemudraia, A., Wiegand, T., Surya, K., Buyukyoruk, M., Cicha, C., Vanderwood, K. K., Wilkinson, R. & Wiedenheft, B. 2020 Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. *Cell Reports Medicine* **1** (6), 100098.

Nicola, M., O'Neill, N., Sohrabi, C., Khan, M., Agha, M. & Agha, R. 2020 Evidence based management guideline for the COVID-19 pandemic. *International Journal of Surgery* **77**, 206–216.

Pan, J., Yang, X., Cai, H. & Mu, B. 2016 Image noise smoothing using a modified Kalman filter. *Neurocomputing* **173**, 1625–1629.

Panchal, G. & Panchal, D. 2015 Solving np hard problems using genetic algorithm. *Transportation* **106**, 6–2.

Park, H. S. & Jun, C. H. 2009 A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* **36** (2), 3336–3341.

Pérez-Cataluña, A., Cuevas-Ferrando, E., Randazzo, W., Falcó, I., Allende, A. & Sánchez, G. 2021 Comparing analytical methods to detect SARS-CoV-2 in wastewater. *Science of the Total Environment* **758**, 143870.

Polasek, W. 1984 Exploring business cycles using running medians. *Computational Statistics & Data Analysis* **2** (1), 51–70.

Press, W. H. & Teukolsky, S. A. 1990 Savitzky-Golay smoothing filters. *Computers in Physics* **4** (6), 669–672.

Price, K., Storn, R. M. & Lampinen, J. A. 2006 *Differential Evolution: A Practical Approach to Global Optimization*. Springer Science & Business Media Berlin Heidelberg:.

Rajagopalan, B. & Lall, U. 1998 Locally weighted polynomial estimation of spatial precipitation. *Journal of Geographic Information and Decision Analysis* **2** (2), 44–51.

Rauch, W., Arabzadeh, R., Grünbacher, D., Insam, H., Markt, R., Scheffknecht, C. & Kreuzinger, N. 2021 Datenbehandlung in der SARS-CoV-2 Abwasserepidemiologie. Korrespondenz Abwasser – in press.

Raudys, A., Lenčiauskas, V. & Malčius, E. 2013 Moving averages for financial data smoothing. In *International Conference on Information and Software Technologies*. Springer, Berlin, Heidelberg, pp. 34–45.

Reinsch, C. H. 1967 Smoothing by spline functions. *Numerische Mathematik* **10** (3), 177–183.

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. 1986 *Akaike Information Criterion Statistics*. D. Reidel, Dordrecht, The Netherlands. Vol. 81(10.5555), 26853.

Samuelsson, O., Björk, A., Zambrano, J. & Carlsson, B. 2017 Gaussian process regression for monitoring and fault detection of wastewater treatment processes. *Water Science and Technology* **75** (12), 2952–2963.

Schutte, J. F., Reinbolt, J. A., Fregly, B. J., Haftka, R. T. & George, A. D. 2004 Parallel global optimization with the particle swarm algorithm. *International Journal for Numerical Methods in Engineering* **61** (13), 2296–2315.

Speckman, P. 1988 Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* **50** (3), 413–436.

Stadler, L. B., Ensor, K., Clark, J. R., Kalvapalle, P., LaTurner, Z. W., Mojica, L., Terwilliger, A. L., Zhuo, Y., Ali, P., Avadhanula, V. & Bertolusso, R. 2020 Wastewater Analysis of SARS-CoV-2 as a Predictive Metric of Positivity Rate for a Major Metropolis. medRχiv. (preprint)

Stone, M. 1978 Cross-validation: a review. *Statistics: A Journal of Theoretical and Applied Statistics* **9** (1), 127–139.

Tusell, F. 2011 Kalman filtering in R. *Journal of Statistical Software* **39** (2), 1–27.

Wand, M. P. 2003 Smoothing and mixed models. *Computational Statistics* **18** (2), 223–249.

Weidhaas, J., Aanderud, Z. T., Roper, D. K., VanDerslice, J., Gaddis, E. B., Ostermiller, J., Hoffman, K., Jamal, R., Heck, P., Zhang, Y. & Torgersen, K. 2021 Correlation of SARS-CoV-2 RNA in wastewater with COVID-19 disease burden in sewersheds. *Science of The Total Environment* **775**, 145790.

Williams, B. M., Durvasula, P. K. & Brown, D. E. 1998 Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record* **1644** (1), 132–141.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C. & Hoelscher, M. 2020 Virological assessment of hospitalized patients with COVID-2019. *Nature* **581** (7809), 465–469.

Wood, S. N., Pya, N. & Säfken, B. 2016 Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111** (516), 1548–1563.

Wu, F., Xiao, A., Zhang, J., Moniz, K., Endo, N., Armas, F., Bonneau, R., Brown, M. A., Bushman, M., Chai, P. R. & Duvallet, C. 2020a SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases. MedrRχiv. (preprint)

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W. L., Armas, F., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N. & Endo, N. 2020b SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. *Msystems* **5** (4), e00614–20.

Yang, L. J., Zhang, B. H. & Ye, X. 2004 Fast Fourier transform and its applications. *Opto-Electronic Engineering* **31**, 1–7.

Zentralanstalt für Meteorologie und Geodynamik ZAMG 2002 *Klimadaten von Österreich 1971–2000*. Available from: http://www.zamg.ac.at/fix/klima/oe71-00/klima2000/klimadaten_oesterreich_1971_frame1.htm

Zhu, Y., Oishi, W., Maruo, C., Saito, M., Chen, R., Kitajima, M. & Sano, D. 2021 Early warning of COVID-19 via wastewater-based epidemiology: potential and bottlenecks. *Science of the Total Environment* **767** (2021), 145124.

First received 20 April 2021; accepted in revised form 16 August 2021. Available online 30 August 2021