Statistica Sinica Preprint No: SS-2022-0413	
Title	Optimally Monitoring a Network of Sewage Manholes in Infectious Disease Surveillance
Manuscript ID	SS-2022-0413
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0413
Complete List of Authors	Leyao Zhang, Yahui Zhang, Chuanwu Xi and Peter X.-K. Song
Corresponding Authors	Peter X.-K. Song
E-mails	pxsong@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

OPTIMALLY MONITORING A NETWORK OF SEWAGE MANHOLES IN INFECTIOUS DISEASE SURVEILLANCE

Leyao Zhang, Yahui Zhang, Chuanwu Xi, Peter X.-K. Song

Department of Biostatistics, University of Michigan

Department of Environmental Health Sciences, University of Michigan

Abstract: The effectiveness of tracking infected cases with the omicron virus has been greatly compromised due to the availability of at-home COVID test kits in many counties. An alternative solution to monitoring contagions of the COVID disease in the population is to survey viral loads from sewage water systems. In a city, hundreds of sewage manholes form a network of candidate sampling sites that are connected to each other in a complex way. Due to the limited resources, it is not viable in practice to sample wastewater from every manhole. The central question of scientific importance is to select those important manholes that are of most relevance to the prediction of confirmed infectious cases in a specific community. In this paper, we develop a supervised learning paradigm of time-series transitional models via the mixed integer programming optimizer to determine important sampling sites to build on a cost-effective monitoring system. We establish the key theoretical guarantee of the selection consistency for the proposed methodology. A novel multi-compartment dynamic model is proposed to simulate viral loads in the wastewater system from the evolution of the pandemic in the population, which is used to evaluate the performance of our proposed model and algorithm. This proposed methodology is illustrated by a real-world data analysis example.

Key words and phrases: COVID-19, GUROBI, L_0 penalization, mixed integer optimization, transitional model.

1. Introduction

With the availability of at-home COVID test kits, it becomes much harder for a public health surveillance system to capture COVID-positive cases in the population. This is because people who test COVID positive at home typically do not report their information of infection to a public disease surveillance program, rather keep the information to themselves. For example, there is no mandatory program for such reporting in the USA. To overcome this significant challenge, scientists have proposed to track COVID viruses from sewage systems as an effective means of monitoring population-level contagions. The current technology allows to extract and quantify three types of RNA targets, including N1 and N2 regions of the nucleocapsid gene and the envelope gene (E), from wastewater samples at labs using primers and probes available at the diagnostic panel assays 2019-

nCoV RUO Kit from the US CDC(CDC, 2021; Corman et al., 2020). Such monitoring techniques are costly in both lab work and manpower, and thus not affordable to run a monitoring program with a large number of sewage manholes. As a result, selecting key monitoring sites from a network of spatially connected sewage manholes to establish a cost-effective disease tracking program is of great importance.

Given the complexity and irregularity of pipeline connections in a given regional sewage infrastructure, it is difficult to use any traditional random field based spatial correlation model such as the Matérn class (Matérn, 2013) or conditional auto-regressive (CAR) model (Besag, 1974) to describe the dependence among manholes. In effect, treating a collection of sewage manholes as a network of nodes appears more appealing in the development of robust methods to optimally determine key monitoring sites in disease surveillance. A subset of important sites may be selected according to their predictability to the daily number of confirmed cases in the area of interest.

In a supervised learning paradigm, selecting key nodes from a network of candidate monitoring manholes is formulated as a constrained optimization involving a binary variable of selecting or not selecting a manhole. This analytic task may be formulated by an integer programming optimization in a time-series transitional prediction model, which is nontrivial and calls for a new method that is both statistically and computationally robust and fast. We propose a mixed integer optimization (MIO) approach and establish the key theoretical guarantee of selection consistency for the proposed method. To evaluate the performance of our statistical estimation, prediction and algorithms, we develop a novel multi-compartment simulation model SEVIR/A that mimics pathways of viral transmission from infected people in the population into a wastewater system that generates sewage samples for surveillance.

Our methodological development is motivated by a sewage monitoring program run by a university in the United States since 2021, where college students contributed primarily to a significant source of infections due to the fact that they lived in highly shared environments (e.g. dorms, classrooms, gyms and dining halls) and had a high likelihood of being asymptomatic due to young ages. This monitoring program selected five spatially scattered dormitories in the college town to periodically (2-3 times per week) collect sewage water samples for the measurements of the N1, N2 and E target genes. To align with daily number of confirmed cases on the campus, these data of gene copy numbers were pre-processed and expanded into daily time series by linear interpolations. Due to time-varying control policies and emerging of new virus variants, to illustrate our methodology using

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

relatively stationary monitoring data, in this paper we chose the period of the 2022 winter semester from January 2, 2022 to April 30, 2022. During this omicron-dominated period of time, individuals who were tested COVID positive were required to report their status to the university health department, so the recorded daily number of confirmed cases on the campus were deemed reliable. When the budget of the program escalated and became costly over time, it was of great interest to figure out if all the five locations were necessary for the monitoring. This calls for developing a statistical methodology to determine optimal spatial sampling schemes to achieve a cost-effective objective in the disease monitoring on the university campus. Indeed, similar questions have been raised in much larger state-level sewage water surveillance networks, so the developed methodology in this paper may be applied, as demonstrated in the simulation studies, to solve such important problems in large-size sewage monitoring networks.

This paper is organized as follows. Section 2 presents formulation of time-series transitional prediction model and parameter estimation via constrained optimization via mixed integer optimization. Section 3 discusses the theoretical properties of the MIO estimator. Section 4 demonstrates the numerical performance of the MIO approach and its competitors via extensive simulation experiments. Section 5 contains the details of the sewage monitoring data analyses. Section 6 gives a few concluding remarks. The technical proof of the MIO selection consistency, software implementation details, additional simulation results and data analysis results are available in the Supplementary Material.

2. Methodology

2.1 Model Formulation with One Monitoring Biomarker

We begin with the case with one monitoring gene, or an RNA biomarker. Consider a time series data containing the daily number of new COVID-19 confirmed cases over a period of T days, denoted by $(y(1), \ldots, y(T))^\top \in \mathbb{N}^{T\times 1}$, and the associated is a daily time series of RNA copy numbers of one target gene (e.g. N1 gene) extracted from sewage water samples, each RNA measurement being collected on one day from a network of m spatially connected manholes, denoted by $\mathbf{G} = \{G_j(1), \ldots, G_j(T)\}_{j=1}^m \in \mathbb{R}^{T\times m}$. We propose to model the connectivity within a network of the sampling manholes (or nodes, or locations) based on a certain network-specific weighting scheme, *say*, $(w_1, \ldots, w_m)^T \in \mathbb{R}^m$. In this paper, w_j is specified as the degree centrality of manhole j, which reflects the fraction of connecting manholes to this node, $j = 1, \ldots, m$. On day t, suggested by scientists, a network-level weighted average number of RNA copies, denoted by $x(t)$, is calculated from the m waste water samples: $x(t) = \frac{\sum_{j=1}^{m} w_j G_j(t)}{\sum_{j=1}^{m} w_j}$. Then, the central question of scientific importance is to establish a linear transitional prediction model for time-series outcome $y(t)$ using P lagged predictors of network-level average RNA biomarkers, $(x(t-1), \ldots, x(t-P))^T$, given by

$$y(t) = \sum_{p=1}^{P} \beta_p x(t-p) + \sum_{q=1}^{Q} \gamma_q y(t-q) + \gamma_0 + \epsilon(t), \qquad (2.1)$$

where γ_0 is the intercept, the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^T \in \mathbb{R}^{P\times 1}$ is of central interest, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_Q)^T \in \mathbb{R}^{Q\times 1}$ is a vector of temporal autoregression coefficients (ARCs), and the error term $\varepsilon(t)$ is sub-Gaussian white noise and independent of historical data of $x(t)$ and $y(t)$. In practice, lags P and Q are both pre-fixed and typically capped by 7 days in the COVID-19 monitoring system. The transitional model is analogous to the time-series autoregression model in the literature, allowing to include lagged predictors on the transitional mechanism of historical time-series outcomes; see Zeger and Qaqish (1988) and Albert and Waclawiw (1998), among others.

Our primary task is to select most relevant manholes in the network of m candidates to predict the number of confirmed cases. To do so, we introduce a set of binary label parameters $\alpha_j, j = 1, \ldots, m$ that help filter out unimportant manholes in the average RNA predictors: $x(t) = \frac{\sum_{j=1}^{m} \alpha_j w_j G_j(t)}{\sum_{j=1}^{m} \alpha_j w_j}$. Of note, parameters $\alpha_j, j = 1, \ldots, m$ take values of 0 or 1, leading to a separation of manholes in the network into two clusters, important $(\alpha_j = 1)$

and unimportant ($\alpha_j = 0$) sampling sites. Using the proposed mixed integer optimization (MIO) algorithm in Section 2.4, we can directly estimate these binary parameters $\alpha_j, j = 1, \ldots, m$ as being 0 or 1, together with the other continuous model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as well as variance σ^2. Given that these model parameters may be binary or continuous, we conduct parameter estimation through an MIO formulation (Bertsimas and Weismantel, 2005; Jünger and Reinelt, 2013), which may be implemented by a commercial MIO solver, GUROBI (Gurobi Optimization, LLC, 2021).

2.2 Estimation via Mixed Integer Optimization

Denote a matrix of parameters by $\boldsymbol{\theta} = (\theta_{p,j}) \in \mathbb{R}^{P \times m}$ with the elements $\theta_{p,j} = \beta_p \alpha_j$, $p = 1, \ldots, P$ and $j = 1, \ldots, m$. MIO formulated as a constrained least squares (LS) estimation in that suitable constraints are specified to control both categorical nature of label parameters α_j and sparsity. The constrained optimization problem is given as follows:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\gamma},\gamma_0} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \gamma_0), \tag{2.2}$$

$$\text{subject to } \boldsymbol{\alpha} \in \{0,1\}^{m\times 1}, \boldsymbol{\theta} \in \mathbb{R}^{P\times m}, \boldsymbol{\beta} \in \mathbb{R}^{P\times 1}, \boldsymbol{\gamma} \in \mathbb{R}^{Q\times 1}, \gamma_0 \in \mathbb{R};$$

$$(1 - \alpha_j)\theta_{p,j} = 0,\ j = 1, \ldots, m,\ p = 1, \ldots, P; \tag{2.3}$$

$$\alpha_j(\theta_{p,j} - \beta_p) = 0,\ j = 1, \ldots, m,\ p = 1, \ldots, P, \tag{2.4}$$

where the LS objective function is

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \gamma_0) =$$
$$\sum_{t=P\vee Q+1}^{T} \left\{ y(t) - \sum_{p=1}^{P} \frac{\sum_{j=1}^{m} \theta_{p,j} w_j G_j(t-p)}{\sum_{j=1}^{m} \alpha_j w_j} - \sum_{q=1}^{Q} \gamma_q y(t-q) - \gamma_0 \right\}^2$$

where $a \vee b = \max(a, b)$. Optimization in (2.2) is to perform an ordinary LS estimation of the model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and γ_0. Each binary parameter α_j flags whether the j-th node is included in the network-level weighted average RNA predictor $x(t)$, and such selection is enforced by two sets of constraints in (2.3) and (2.4). More precisely, when $\alpha_j = 0$, constraint (2.3) necessitates $\theta_{p,j} = 0$; when $\alpha_j = 1$, constraint (2.4) compels $\theta_{p,j} = \beta_p$. As a results, all $\theta_{p,j}$'s with $\alpha_j = 1$ are fused into a common parameter β_p, namely the effect of a lag-p average RNA predictor $x(t-p)$ on outcome $y(t)$. In other words, among those selected nodes with $\alpha_j = 1$, we have $\theta_{p,j} = \beta_p \alpha_j = \beta_p$. After obtaining the constrained LS estimation above, we estimate the parameter σ^2 using the sample variance of the residuals, $e(t) = y(t) - \sum_{p=1}^{P} \frac{\sum_{j=1}^{m} \hat{\theta}_{p,j} w_j G_j(t-p)}{\sum_{j=1}^{m} w_j \hat{\alpha}_j} - \sum_{q=1}^{Q} \hat{\gamma}_q y(t-q) - \hat{\gamma}_0, t = \max(P, Q) + 1, \ldots, T$.

Denote the oracle estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\theta}}^{(ol)}$ and $\hat{\boldsymbol{\gamma}}^{(ol)}$; they would be obtained directly from the unconstrained LS estimation when the true labels of important monitoring sites were known, so calculating the network-level RNA average $x(t)$ were straightforward. Both theory and numerical performance of the least squares estimation in the oracle transitional model has been extensively validated in the literature (Brockwell and Davis, 2002). Technically, given the set of true labels $\boldsymbol{\alpha}_0 \in \{0,1\}^m$, the oracle solutions are obtained by

$$(\hat{\boldsymbol{\theta}}^{(ol)\top}, \hat{\boldsymbol{\gamma}}^{(ol)\top})^\top := \underset{\boldsymbol{\theta}=\boldsymbol{\beta}\boldsymbol{\alpha}_0^\top, \boldsymbol{\gamma}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \gamma_0). \tag{2.5}$$

When the true labels are unknown, finding solutions close to the oracle estimates may be carried out by the all-subsets regression (ASR) that produces the L_0 penalized estimates with no homogeneity pursuit (Ke et al., 2015); that is, constrains (2.3) and (2.4) are not involved in the LS estimation to fuse non-zero estimates into one cluster. One state-of-the-art ASR method is the Adaptive Best-Subset Selection (ABESS) algorithm (Zhu et al., 2020), which will be compared with our proposed MIO method in numerical examples in Section 4. When the number of nodes in the network is bigger than 20, ASR becomes too computationally burdensome to enumerate all possible subset configurations, thus becomes unfeasible in practice. The MIO formulation provides an appealing alternative to greatly extend the capacity of ASR with parameter fusion as part of the solution.

2.3 Model Extensions with Multiple Monitoring Biomarkers

In practice, three types of RNA biomarkers (e.g., N1, N2 and E genes) are typically measured for the COVID-19 disease monitoring, and using them requires an extension of the single-gene prediction model (2.1). Consequently, the MIO formulation in (2.2)-(2.4) will be extended to determine a subset of important sites based on multiple monitoring biomarkers.

Suppose that $K(K \geq 1)$ types of genetic markers, $G_j^k(t), k = 1, \ldots, K$, have been quantified from sewage water samples collected in a network of m sites on the the daily basis over a period of T days. An extension of the linear transitional prediction model (2.1) takes the following form:

$$y(t) = \sum_{p=1}^{P} \sum_{k=1}^{K} \beta_{p,k} x_k(t-p) + \sum_{q=1}^{Q} \gamma_q y(t-q) + \gamma_0 + \epsilon(t), \qquad (2.6)$$

where $x_k(t) = \frac{\sum_{j=1}^{m} \alpha_j w_j G_j^k(t)}{\sum_{j=1}^{m} \alpha_j w_j}$ is the k-th network-level weighted average of RNA copies. Because K is a fixed dimension, this extended model (2.6) shares similar theoretical results of selection consistency established for the one-biomarker model (2.1) in Section 3.

2.4 Implementation

We apply the GUROBI Optimizer (Gurobi Optimization, LLC, 2021) to solve the MIO defined by equations (2.2) - (2.4). The numerical task is to minimize the LS loss function in (2.2) under the constrains given by (2.3)-(2.4) for the binary cluster label parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$. Both constraints are referred to as Type 1 Special-Ordered Set (SOS) constraints (Beale et al., 1969) in that at most one variable in a set takes a nonzero value. It is worth reiterating that by the MIO approach, the cluster labels of being relevant (1) or irrelevant (0) are obtained via a joint operation of optimization together with all the other continuous model parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and γ_0. Algorithm 1 in the Supplementary Material lists the key steps required in the pseudo code that outputs the MIO solution to the constrained optimization problem. In both simulation studies and data analyses, we use GUROBI version 9.5.2 in all numerical calculations.

3. Theoretical Guarantees

In this section we establish the key large-sample property for the MIO estimator of the parameter $\boldsymbol{\theta}$ obtained by the constrained optimization given in (2.2)-(2.4). For simplicity, we consider the case of the one-biomarker transitional model in (2.1). Under some mild regularity conditions, Theorem 1 below presents the selection consistency, which is the theoretical basis for the application of the proposed MIO method.

Condition 1. *(Sub-Gaussian white noise) The error terms $\epsilon(t)$ in model*

(2.1) are independent sub-Gaussian with mean zero and Ψ_2-norm bounded by σ (Vershynin, 2018).

Not every MIO problem is solvable. To quantify the difficulty of identifying the set of true labels, following Shen et al. (2013), we define a grouping sensitivity measure with respect to incorrect subset selection:

$$\begin{aligned} c_{\min} &\equiv c_{\min}(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \mathbf{G}_w, \mathbf{y}_0) \\ &= \min_{\substack{\boldsymbol{\alpha}\in\{0,1\}^m, \boldsymbol{\alpha}\neq\boldsymbol{\alpha}_0 \\ \boldsymbol{\theta}=\beta\boldsymbol{\alpha}, \boldsymbol{\gamma}}} \frac{\|\mathbf{G}_w(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\mathbf{Z}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\|_2^2}{(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)}, \end{aligned} \tag{3.1}$$

where $d(\boldsymbol{\alpha}, \boldsymbol{\alpha}') := \mathbf{1}^\top\boldsymbol{\alpha} + \mathbf{1}^\top\boldsymbol{\alpha}' - 2\boldsymbol{\alpha}^\top\boldsymbol{\alpha}'$ quantifies the distance between two sets of labels $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, where $\mathbf{1}$ is an m-dimensional vector whose entries are all 1. Clearly, d is the minimum number of entries that need to be altered to reach $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$. Term $c_{\min}$ in (3.1) may be regarded as the minimum increase of MSE per falsely classified sampling node. A small value of $c_{\min}$ suggests that the MSE is insensitive to any false subset selection, thus makes an algorithm difficult to select the true labels. In other words, the smaller the $c_{\min}$ is, the more difficult the optimization problem is.

Theorem 1. *Let $\hat{\boldsymbol{\theta}}$ be the MIO estimator obtained from equations (2.2)-(2.4) and $\hat{\boldsymbol{\theta}}^{(ol)}$ is the oracle estimator given in equation (2.5). Under the regularity condition of sub-Gaussian white noise above, if $c_{\min} > \frac{40\sigma^2}{3\{T-\max(P,Q)\}}\log(m)$, then $\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{(ol)}) \to 0$ as $T, m \to \infty$. This implies that when $c_{\min} >$*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

$\frac{40\sigma^2}{3\{T-\max(P,Q)\}}\log(m)$, *the MIO estimator* $\hat{\boldsymbol{\theta}}$ *consistently reconstructs the oracle estimator* $\hat{\boldsymbol{\theta}}^{(ol)}$, *and* $\mathbb{P}(\hat{\boldsymbol{\alpha}} \neq \hat{\boldsymbol{\alpha}}^{(ol)}) \to 0$ *as* $T, m \to \infty$.

Theorem 1 implies that the MIO estimator $\hat{\boldsymbol{\theta}}$ consistently reconstructs the oracle estimator $\hat{\boldsymbol{\theta}}^{(ol)}$ in equation (2.5) as the sample sizes $T, m \to \infty$. The proof of Theorem 1 is given in the Supplementary Material Section 1.

The theoretical benefit of node selection via the above fusion procedure lies in its capacity of addressing the challenge of selecting manholes with weak viral signals. The traditional selection analytic via p-values is known to perform poorly in identifying weak signals, often yielding undesirable model fit. Our numerical experiences suggest that such aggregated features derived from the fusion technique gain increased tolerance for spurious signals, enhancing overall model robustness.

4. Simulation Experiments

We conduct simulations to assess the finite-sample performance of the proposed MIO method to determine important sampling locations from a network in disease surveillance. We compare our MIO Algorithm with existing methods, LASSO (Tibshirani, 1996) and ABESS (Zhu et al., 2020). We employ the Python package ‘sklearn.linear_model.LassoCV’ (Pedregosa et al., 2011) for LASSO, tuning sparsity through 20-fold cross-validation. A manholes is labeled as "important" if its viral signal has a nonzero LASSO association estimate; otherwise, it is labeled "unimportant". Such a procedure is applied to ABESS in the operation of the Python package 'abess'.

4.1 Transitional model

We simulate 1000 datasets from the following transitional model (4.1), each containing a daily time series of COVID-19 confirmed cases, $y(t)$, and a daily time series of RNA copies, $G_j(t)$, from a network of m manholes over a period of 365 days that leads to the network-level average predictor $x(t) = \frac{\sum_{j=1}^m \alpha_j w_j G_j(t-1)}{\sum_{j=1}^m \alpha_j w_j}$. The transitional model takes the following form:

$$y(t) = \beta x(t) + \gamma_1 y(t-1) + \gamma_0 + \varepsilon(t), t = 2, \dots, 365, \tag{4.1}$$

where the number of manholes is set at $m \in \{20, 50, 100\}$, the intercept $\gamma_0 = 0$, the autoregression coefficient $\gamma_1 = 0.3$, and the errors $\varepsilon(t) \overset{iid}{\sim} N(0,1), t = 2, \dots, 365$, as well as the initial condition for $y(1) = 0$.

We consider three scenarios of important nodes to the prediction of confirmed cases, specified respectively as 20%, 50% and 80% of the nodes in the network of sewage manholes. The network topology is generated by the stochastic block model (Holland et al., 1983; Zhao et al., 2012; Jin et al., 2023), in which the nodes belong to three clusters with sizes $\frac{3}{5}m$, $\frac{1}{5}m$, and $\frac{1}{5}m$. The within-cluster edge probability is 0.5, while between-cluster probabilities are set as follows: 0.25 for clusters 1 and 2, 0.1 for clusters 2 and 3, and 0.1 for clusters 1 and 3. The resulting degree centrality is then determined as node-specific weight w_j for the network. Figure 2 displays three simulated networks of 20, 50, 100 nodes, respectively.

4.2 Infectious disease model for viral RNA copies

To simulate RNA copies in the sewage system, we first simulate the number of infections in the population through a multi-compartment SEVIR/A model, from which we then simulate the RNA concentration data in the wastewater through a viral shedding model.

Extended from the eSAIR model (Zhou et al., 2020), as shown in Figure 1, SEVIR/A is a Susceptible (S)-Antibody (A)-Infectious (I)-Recovered (R) model for the population dynamics of COVID-19 pandemics useful to simulate the daily number of new infections in the population. $S(t)$, $E(t)$, $V(t)$, $I(t)$ and $R(t)$ denote, respectively, the number of individuals in one compartment at a given day t, starting at $t = 0$. The SEVIR/A model is specified by the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dS(t)}{dt} &= r_{R/AS}R(t) - \frac{r_{SE}\pi(t)S(t)I(t)}{N} - r_{SV}S(t),\\
\frac{dV(t)}{dt} &= r_{SV}S(t) - r_{VR/A}V(t),\\
\frac{dE(t)}{dt} &= \frac{r_{SE}\pi(t)S(t)I(t)}{N} - r_{EI}E(t),\\
\frac{dI(t)}{dt} &= r_{EI}E(t) - r_{IR/A}I(t),\\
\frac{dR(t)}{dt} &= r_{IR/A}I(t) + r_{VR/A}V(t) - r_{R/AS}R(t),
\end{aligned}$$

where N denotes the total population size. In the simulation study, N is set at 40000 that mimics a university campus population. In addition, according to Wamalwa and Tonnang (2022) and McMahan et al. (2021), we set the disease transmission rate $r_{SE} = 0.26$, the vaccination rate $r_{SV} = 0.00036$, and the rate of contagion $r_{EI} = 0.2$. In the meanwhile, given that the average recovery period is about two weeks and the recurrent susceptibility period is about three months, we set the rate of noninfectious $r_{IR/A} = 0.1$, the rate of immunization $r_{VR/A} = 0.07$ and the rate of recurrent susceptibility $r_{R/AS} = 0.01$. In addition, following Wang et al. (2020), we multiply the transmission rate β with a time-varying transmission rate modifier $\pi(t) = 1 + \sigma_1 \cos(2\pi\frac{t-\zeta_1}{365}) + \sigma_2 \cos(2\pi\frac{t-\zeta_2}{365}) + \sigma_3 \cos(2\pi\frac{t-\zeta_3}{365}), t = 1, \ldots, 365$, which incorporates seasonality in the model. We set $\sigma_1, \sigma_2, \sigma_3$ at 0.1, 0.6, 0.2 and set $\zeta_1, \zeta_2, \zeta_3$ at 20, 150, 250 to mimic the three peaks in the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

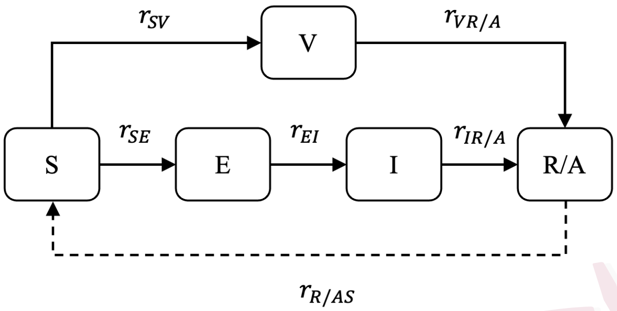

Figure 1: The SEVIR/A model is used to simulate the dynamics of the COVID-19 pandemic in the population.

observed daily number of confirmed cases in the motivating data. Using the R package `deSolve`, we obtain the daily number of new infections from the above SEVIR/A model.

Viral shedding from infectious individuals causes the disease spread in the population, part of which can be measured from the sewage system. Following Cavany et al. (2022), the viral RNA copies, $G_j(t)$, on day t in the wastewater at site j are generated from a negative binomial (NB) distribution with the size $r = 0.393$ and the mean viral RNA concentration given by $C_m(t) = c_0 \sum_p \Delta I(t-p)\sigma_i(p)$, where c_0 is the scaling constant to capture the magnitude of shedding (set at 2 in the simulation), $\Delta I(t)$ is the daily number of new infections given from the SEVIR/A model, $\sigma_i(p)$ is the individual shedding distribution on p days after infection, which is found to

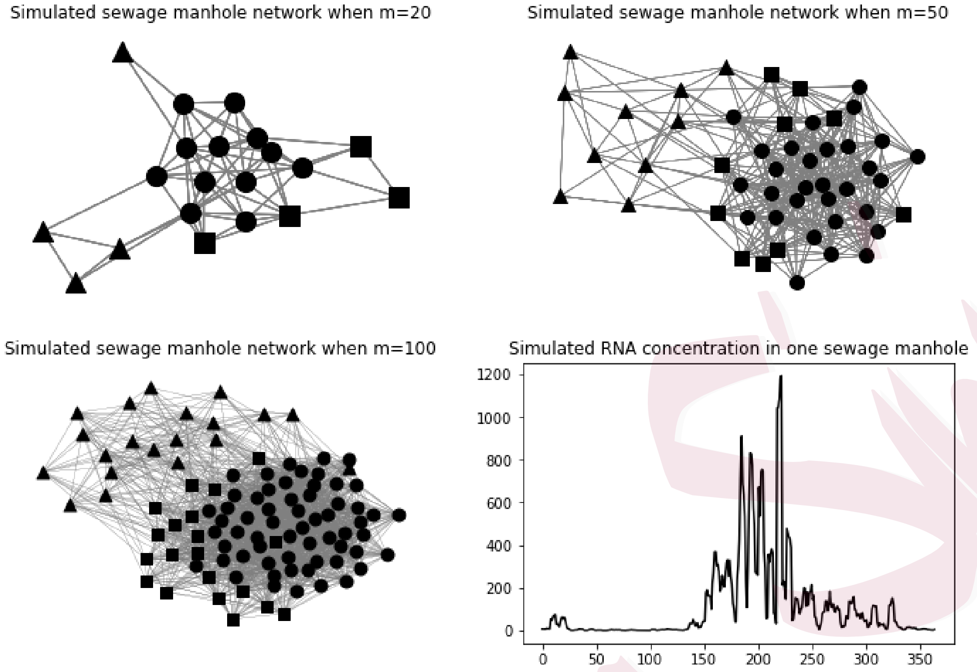

Figure 2: Three simulated networks of $m = 20, 50$ and 100 manholes, respective, and a simulated time series of RNA gene copies at one manhole over 365 days.

follow approximately a Gamma distribution with the shape 2.05 and rate 0.169 (Cavany et al., 2022). See Panel (2,2) in Figure 2 for a simulated time series of the RNA copies in one sampling manhole over 365 days.

4.3 Results

Selection Accuracy. We use normalized mutual information (NMI) to assess the performance of selecting important nodes, and a higher NMI value indicates a better node selection (Lancichinetti et al., 2009) (see Supplementary Material Section 2.1 for a detailed definition of NMI). Its upper limit (i.e. $NMI = 1$) reflects a perfect grouping with no labeling error while its lower limit (i.e. $NMI = 0$) signifies a complete clustering failure. Table 1 demonstrates the performance of subset classification into important versus unimportant nodes by our proposed MIO method and two existing methods of popularity, LASSO and ABESS. It is evident that the MIO method has exhibited a superior performance regardless of the network size and sparsity level (i.e. the number of important nodes). The MIO Algorithm in all cases except two delivers above 96% selection accuracy, even for the cases with a large number of manholes ($m = 100$). The ABESS algorithm has shown a comparable performance in the cases of high sparsity (i.e. a small number of important manholes) and the strong signal (or large β). However, when the number of important manholes increases, the selection accuracy deteriorates significantly. LASSO performs poorly due possibly to dependencies among nodes, and thus is not recommended for the estimation of cluster labels. Results in Table 1 clearly unveils that the difficulty of the underlying optimization increases with higher complexity of network and low signal-to-noise ratio. That is, smaller effect size β, bigger network size m or lower sparsity level s presents clear challenges in achieving desirable

Table 1: Average Normalized Mutual Information (NMI) over 1000 replicates by three methods for the selection accuracy, where m is the network size, s is the number of important manholes, and β is the signal strength.

$m = 20$

	$s = 4$		$s = 10$		$s = 16$	
Method	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
MIO	0.975	0.972	0.987	0.982	0.994	0.990
LASSO	0.130	0.129	0.057	0.087	0.046	0.059
ABESS	0.948	0.942	0.927	0.886	0.900	0.839

$m = 50$

	$s = 10$		$s = 25$		$s = 40$	
Method	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
MIO	0.998	0.962	0.994	0.929	0.997	0.964
LASSO	0.077	0.065	0.029	0.042	0.026	0.022
ABESS	0.985	0.854	0.855	0.370	0.743	0.217

$m = 100$

	$s = 20$		$s = 50$		$s = 80$	
Method	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$	$\beta = 0.8$	$\beta = 0.4$
MIO	0.997	0.962	0.997	0.874	0.998	0.961
LASSO	0.030	0.034	0.016	0.015	0.000	0.017
ABESS	0.935	0.510	0.228	0.095	0.069	0.050

selection accuracy. Refer to Tables 1 and 2 in the Supplement Material for comparisons of selection sensitivity and specificity among these methods.

Parameter Estimation. Performance of estimation is evaluated by average absolute bias (AAB) and empirical standard error (ESE) for the model parameter β over 1000 replications. In the comparison, the oracle method refers to the estimation of β obtained under the known true labels.

Table 2 shows that the MIO performs very well, and even in the cases in which the selection accuracy is not perfect, the size of estimation bias is comparable to that in the cases with the oracle as well as those with the perfect selection. Under a mild departure from the perfect selection (say 10% or less), the MIO exhibits desirable estimation accuracy of the key parameter β. In contrast, ABESS and LASSO have difficulty in handling low sparsity, where estimation biases are significantly higher than those given by the MIO. Table 3 in the Supplementary Material displays AAB and ESE for MIO's estimation of parameter $\boldsymbol{\gamma}$, indicating its superiority over existing methods in terms of estimation bias.

In addition to the above scenario of randomly generated network topology, we also assess the performance of the MIO method in static networks with a fixed number of important nodes and their connectivity. Tables 4 and 5 in the Supplementary Material reaffirm the MIO method's high selection accuracy and stability in the scenario of fixed network topology.

Furthermore, we validate the stability of GUROBI software for obtaining optimal solutions. Refer to Table 6 in the Supplementary Material.

Table 2: Average absolute bias (empirical standard error) of $\hat{\beta}$ over 1000 replicates by four methods, where Oracle uses the true group labels while LASSO and ABESS estimate group labels by nonzero estimates.

	$m = 20$					
	$s = 4$		$s = 10$		$s = 16$	
Method	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
Oracle	0.001 (0.001)	0.001 (0.001)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)
MIO	0.001 (0.001)	0.001 (0.001)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)
LASSO	0.130 (0.007)	0.058 (0.004)	0.103 (0.004)	0.038 (0.007)	0.377 (0.003)	0.077 (0.006)
ABESS	0.007 (0.002)	0.004 (0.001)	0.011 (0.002)	0.007 (0.002)	0.032 (0.003)	0.017 (0.002)

	$m = 50$					
	$s = 10$		$s = 25$		$s = 40$	
Method	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
Oracle	0.002 (0.002)	0.001 (0.002)	0.002 (0.003)	0.002 (0.003)	0.003 (0.004)	0.003 (0.004)
MIO	0.002 (0.002)	0.002 (0.002)	0.002 (0.003)	0.003 (0.003)	0.003 (0.004)	0.003 (0.004)
LASSO	0.050 (0.005)	0.023 (0.003)	0.092 (0.008)	0.031 (0.005)	0.388 (0.003)	0.080 (0.007)
ABESS	0.002 (0.002)	0.003 (0.002)	0.006 (0.003)	0.024 (0.003)	0.011 (0.004)	0.053 (0.004)

	$m = 100$					
	$s = 20$		$s = 50$		$s = 80$	
Method	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$	$\beta = 0.8$	$\beta = 0.4$
Oracle	0.002 (0.003)	0.002 (0.003)	0.003 (0.004)	0.003 (0.004)	0.004 (0.005)	0.004 (0.005)
MIO	0.002 (0.003)	0.002 (0.003)	0.003 (0.004)	0.004 (0.004)	0.004 (0.005)	0.004 (0.005)
LASSO	0.109 (0.010)	0.044 (0.006)	0.392 (0.003)	0.082 (0.010)	0.800 (0.000)	0.392 (0.002)
ABESS	0.008 (0.003)	0.021 (0.003)	0.131 (0.009)	0.100 (0.004)	0.330 (0.015)	0.201 (0.007)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

Run Time. Table 3 compares the median run time of the MIO to the two existing methods. All computations are parallelized in the Advanced Research Computing facility at the University of Michigan where each task is operated by a CPU with 4K MB memory. Although the MIO tends to be the slowest it can complete most of the computational jobs within 5 minutes regardless of network size m. This computational cost is practically acceptable, given that the MIO clearly gained much higher numerical quality in both selection accuracy and estimation accuracy.

5. Data Application

We apply the proposed method to analyze the motivating data collected from a sewage monitoring study. In this study, N1 and N2 gene copies were extracted and quantified by a biological laboratory from sewage water samples collected at five on-campus dormitories during January 1 to April 30, 2022 when the omicron variant dominated the pandemic. The outcome of interest in our analysis is daily time series of on-campus confirmed COVID-19 cases, which was collected daily by the university health department. The left panel of Figure 3 depicts the network of manholes exclusively connected to five manholes, labeled as A, B, C and D (i.e. the red dots), which are chosen and approved for the study by the university to collect wastew-

Table 3: Median run time (in seconds) over 1000 replicates.

			$m = 20$			
	$s = 4$		$s = 10$		$s = 16$	
	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
MIO	2.547	2.420	2.201	2.172	2.012	1.999
LASSO	0.168	0.176	0.173	0.190	0.141	0.183
ABESS	0.004	0.004	0.004	0.004	0.004	0.004
			$m = 50$			
	$s = 10$		$s = 25$		$s = 40$	
	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$
MIO	8.991	8.625	6.290	6.444	5.943	6.038
LASSO	0.327	0.388	0.338	0.409	0.157	0.388
ABESS	0.018	0.018	0.018	0.018	0.019	0.019
			$m = 100$			
	$s = 20$		$s = 50$		$s = 80$	
	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.4$	$\beta = 0.2$	$\beta = 0.8$	$\beta = 0.4$
MIO	35.616	34.714	30.746	300.751	46.780	124.267
LASSO	0.649	0.801	0.189	0.890	0.169	0.190
ABESS	0.016	0.016	0.016	0.017	0.017	0.017

ater samples (their real names and locations are anonymous due to the data security consideration). The grey dots are other manholes associated with administration, classrooms and residents buildings. The right panel of Figure 3 shows the serial boxplots of weekly log-transformed on-campus confirmed cases (the daily data are not shown due to the confidentiality).

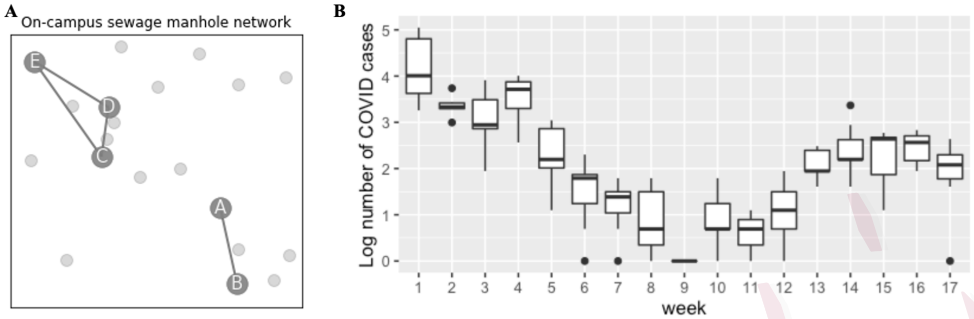

Figure 3: (A)On-campus sewage manhole network and (B) the weekly distributions of log-transformed on-campus confirmed cases.

Applying the MIO approach, we aim to derive a subset of important monitoring sites to make the study more cost-effective. As part of validation for the MIO solution, we compare it with an all-subsets regression (ASR) solution, which exhaustively searches for the best subset of sampling nodes for the prediction of the infection case counts. Here we have a 5-dimensional vector of label parameters $(\alpha_1, \dots, \alpha_5)^\top \in \{0,1\}^5$ that gives rise to $2^5 = 32$ possible configurations of subsets. Thus, after fitting 32 versions of the linear model (5.1) below, we obtain the global optimal solution in terms of the smallest adjusted R^2. Due to the small size network in this application, we can fortunately obtain such gold standard reference to assess the performance of the MIO method. Of note, when the network size increases to 20 or bigger, the exhaustive ASR method becomes computationally prohibited, and thus obtaining the global optimal solution is no longer numerically viable.

The transitional prediction model for the daily time series of log-transformed on-campus confirmed cases takes the following form with both daily time series predictors of RNA copies from N1 gene and N2 gene:

$$y(t) = \sum_{p=1}^{7}\sum_{k=1}^{2}\beta_{p,k}x_k(t-p) + \sum_{q=1}^{7}\gamma_q y(t-q) + \gamma_0 + \epsilon(t), t = 8, \ldots, 119. \quad (5.1)$$

To yield meaningful interpretations for both parameter estimates and prediction, we add the additional constraints of positivity on the model parameters, $\beta_{p,k} \geq 0,\ p = 1, \ldots, 7,\ k = 1, 2$ and $\gamma_q \geq 0,\ q = 1, \ldots, 7$. Consequently, these coefficients can be estimated as exact zeros. The normality assumption of the error term $\epsilon(t)$ is checked by the Q-Q plot of the residuals; see Figure 1 in the Supplementary Material.

The MIO solution selects manholes connected three dormitories A, D and E that are spatially spanned to gain maximal coverage. The MIO estimates are listed in Table 4, where the lag-p network-level average RNA predictor is computed from a subset of selected sites. In Figure 3, A represents the southern campus, while E is located at the edge of the north campus and D is at the heart of the north campus near a large complex of classroom buildings. In fact, the MIO solution is globally optimal as it coincides with the gold standard obtained by the exhaustive ASR method.

Thus, as far as the prediction of on-campus COVID-19 confirmed cases concerns, with a high confidence we recommend reducing the monitoring program involving 5 manholes to a reduced program with three manholes connecting to dormitories A, D and E. This potentially leads to a 40% saving of costs and manpower.

For the purpose of comparison, we also fit model (5.1) with no use of subset selection, namely $\alpha_j = 1, j = 1, \ldots, 5$. The model built upon the MIO subset solution with the network-level average biomarkers has an adjusted R^2 equal to $R_a^2 = 0.736$, which is higher than the model with no selection of manholes, $R_a^2 = 0.695$. This subset selection implemented in the MIO approach does help improving goodness-of-fit over the naive method using all 5 sampling nodes in the network.

6. Concluding Remarks

In this paper we developed a supervised learning framework that enables to select a subset of important monitoring nodes from a network of spatially connected sewage manholes. This resulting subset is intended to not only improve prediction of daily COVID-19 confirmed cases but also to make the monitoring program more cost-effective. Our methodological development utilized the mixed integer optimization (MIO), which is implemented by the

Table 4: Regularized point estimates of the transitional model parameters in the sewage monitoring study, including $\beta_{p,1}$ for lagged N1 gene predictors, $\beta_{p,2}$ for lagged N2 gene predictors, and γ_q for ARCs over 7 days.

Lag	N1	est.$\times 10^{-4}$	N2	est.$\times 10^{-4}$	ARC	est.
Day 0	-	-	-	-	γ_0	0.17
Day 1	$\beta_{1,1}$	0.00	$\beta_{1,2}$	0.00	γ_1	0.50
Day 2	$\beta_{2,1}$	6.69	$\beta_{2,2}$	0.00	γ_2	0.12
Day 3	$\beta_{3,1}$	0.00	$\beta_{3,2}$	0.00	γ_3	0.00
Day 4	$\beta_{4,1}$	0.00	$\beta_{4,2}$	0.00	γ_4	0.00
Day 5	$\beta_{5,1}$	0.00	$\beta_{5,2}$	0.60	γ_5	0.00
Day 6	$\beta_{6,1}$	0.00	$\beta_{6,2}$	0.47	γ_6	0.23
Day 7	$\beta_{7,1}$	0.00	$\beta_{7,2}$	0.00	γ_7	0.00

commercial software GUROBI. Under some mild regularity conditions, the theoretical guarantee of selection consistency was established. Through extensive numerical experiments, we demonstrated the desirable performances of the proposed MIO approach that outperformed some popular methods in the current literature, including LASSO and ABESS.

Computational capacity is a clear challenge for the MIO approach. While the GUROBI solver performs adequately for networks with up to 100 nodes, it can become computationally slow or even fails to work as the network size increases. One solution is to develop more powerful algorithms and software such as those recently developed, for example, in Bertsimas et al. (2016). Another solution is to take a sure screening step reduce network dimensionality to a level manageable by the GUROBI solver.

The proposed methodology may be applied to analyze a much larger COVID-19 virus monitoring network spanning the entire state of Michigan, with data collected from more than 200 manholes and wastewater treatment plants. The related results will be published elsewhere once the statewide data use agreement get approved in the future. Note that we have effectively structured the simulation experiment to emulate the envisioned scale for statewide implementation to show the technical feasibility.

Some useful extensions from the current methodology include (i) to allow time-domain selection of specific days for sample collection, instead of current daily sampling, which certainly leads to more cost savings; (ii) to extend the ordinary least squares objective function to a likelihood objective, so that generalized linear models such as the Poisson log-linear model can be applied to analyze time series of counts; and (iii) to improve the flexibility of network-level summary biomarker, so that the model can accommodate and enjoy finer network topology into the fusion of nodes, such as multiple sub-community level summary markers.

Supplementary Material

This document of Supplementary Materials includes (A) the technical proof of the selection consistency of the MIO estimators; (B) software implementation details; (C) additional simulation results and (D) additional data analysis results.

Acknowledgments

The authors are grateful to both Associate Editor and an anonymous reviewer for their constructive comments that helped improve the paper greatly. This research was supported primarily by a grant E20234148 from the Michigan Department of Health and Human Services, and in part by a NSF grant DMS2113564 to Song.

References

Albert, P. and M. Waclawiw (1998). A two-state Markov chain for heterogeneous transitional data: A quasi-likelihood approach. *Statistics in Medicine 17*, 1481–1493.

Beale, E., , and J. A. Tomlin (1969). Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables. *Operational Research 69*(99), 447–454.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics 44*(2), 813–852.

Bertsimas, D. and R. Weismantel (2005). *Optimization over Integers*. Dynamic Ideas.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(2), 192–225.

Brockwell, P. J. and R. A. Davis (2002). *Introduction to time series and forecasting*. Springer.

Cavany, S., A. Bivins, Z. Wu, D. North, K. Bibby, and T. Perkins (2022). Inferring SARS-CoV-2 RNA shedding into wastewater relative to the time of infection. *Epidemiology and Infection 150*(E21).

CDC (2021). CDC 2019-novel coronavirus (2019-nCoV) real-time RT-PCR diagnostic panel. https://www.fda.gov/media/134922/download.

Corman, V., O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. Chu, and et al. (2020). Detection of 2019 novel coronavirus (2019-ncov) by real-time RT-PCR. *Eurosurveillance 25*, 2000045.

Gurobi Optimization, LLC (2021). Gurobi Optimizer Reference Manual. https://www.gurobi.com.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Jin, J., Z. T. Ke, S. Luo, and M. Wang (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association 118*(543), 2101–2116.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0413

Jünger, M. and G. Reinelt (2013). *Facets of Combinatorial Optimization*. Springer.

Ke, Z. T., J. Fan, and Y. Wu (2015). Homogeneity pursuit. *Journal of the American Statistical Association 110*(509), 175–194.

Lancichinetti, A., S. Fortunato, and J. Kertész (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics 11*(3), 033015.

Matérn, B. (2013). *Spatial variation*, Volume 36. Springer Science & Business Media.

McMahan, C. S., S. Self, L. Rennert, C. Kalbaugh, D. Kriebel, D. Graves, and et al. (2021). COVID-19 wastewater epidemiology: a model to estimate infected populations. *The Lancet Planetary Health 5*(12).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Shen, X., W. Pan, Y. Zhu, and H. Zhou (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics 65*(5), 807–832.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Wamalwa, M. and H. E. Tonnang (2022). Using outbreak data to estimate the dynamic covid-19 landscape in eastern africa. *BMC Infectious Diseases 22*(1), 531.

Wang, L., Y. Zhou, J. He, B. Zhu, F. Wang, L. Tang, and et al. (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in china. *Journal of Data Science 18*(3), 409–432.

Zeger, S. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics 44*, 1019–1031.

Zhao, Y., E. Levina, and J. Zhu (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics 40*(4), 2266–2292.

Zhou, Y., L. Wang, L. Zhang, L. Shi, K. Yang, J. He, and et al. (2020). A Spatiotemporal Epidemiological Prediction Model to Inform County-Level COVID-19 Risk in the United States. *Harvard Data Science Review* (Special Issue 1), 1–33. https://hdsr.mitpress.mit.edu/pub/qqg19a0r.

Zhu, J., C. Wen, J. Zhu, H. Zhang, and X. Wang (2020). A polynomial algorithm for best-subset selection problem. *PNAS 117*(52), 33117–33123.

Leyao Zhang, Department of Biostatistics, University of Michigan

E-mail: (leyaozh@umich.edu)

Yahui Zhang, Department of Biostatistics, University of Michigan

E-mail: (yahuiz@umich.edu)

Chuanwu Xi, Department of Environmental Health Sciences, University of Michigan

E-mail: (cxi@umich.edu)

Peter X.K. Song, Department of Biostatistics, University of Michigan

E-mail: (pxsong@umich.edu)