A Transfer Learning-Based Multivariate Control Chart for Dengue Surveillance in Hong Kong

ZEZHONG WANG AND INEZ MARIA ZWETSLOOT

ABSTRACT Dengue is a severe mosquito-borne epidemic disease. There is no effective vaccine for dengue, so a real-time surveillance system becomes crucial to detect dengue outbreaks. Control charts have been widely used as efficient tools to identify changes in health-related data. In Hong Kong, the environmental protection department uses the area ovitrap index to survey monthly the number of mosquitoes in different areas. Some areas have limited historic area ovitrap index records since the survey started only recently. Parameter estimation for designing control charts is challenging with little historic data. This paper proposes a transfer learning-based estimator to increase parameter estimation accuracy for areas with limited historic data points. We study the questions on what and how to transfer useful knowledge from related source data. A multivariate control chart based on transfer-learned parameters is developed for online monitoring. A real example of dengue surveillance demonstrates its effectiveness in application.

INDEX TERMS Health surveillance, transfer learning, statistical process monitoring (SPM), multivariate control chart, dengue surveillance.

I. INTRODUCTION

Public health surveillance is a health-related data collection and analysis system for detecting outbreaks of diseases. It enables public health institutions to monitor the current health status and prepare for potential health emergencies. The recent global outbreak of coronavirus has highlighted the importance of developing more efficient and effective health surveillance techniques to enhance the public health system’s readiness to deal with unexpected shocks. The surveillance effort often involves retrospective studies and prospective studies. In the retrospective studies, historical data are fixed and used to find the prevalence or patterns of disease [2], and to provide an overview of health forecasting methods [15]. In the prospective studies, the principal object is timely detection of outbreaks and enabling effective control measures.

References [10] and [16] outlined technical issues and challenges in health surveillance. Prospective health surveillance techniques have faced challenges in development and implementation, as noted in reviews by [2], [5], [9], and [14].

As discussed by [3], control charts can meet the requirement of health surveillance. Reference [26] reviews aberration detection methods in public health surveillance, and concludes that control charts are commonly used in practice. Reference [4] proposed a one-sided MEWMA control chart to monitor disease rates in several regions. Reference [12] develops a real-time surveillance method for detecting emerging disease outbreaks via both daily residuals and AR model-based detrended residuals. Reference [29] proposes a flexible spatiotemporal monitoring method for the Florida influenza-like illness data. The method consists of building a baseline model and then monitoring the observed disease incidence rates. Reference [40] proposed a distribution-free control chart based on change-point analysis for detecting changes in the location of univariate influenza-like illness rate data. Reference [8] compared the performance of the scan statistics method, CUSUM chart, and EWMA chart in detecting disease rates. For a recent review of using control charts in the public health domain, see reference [39].

Dengue is a viral disease transmitted to humans through the bites of infected mosquitoes, and its global incidence has been increasing steadily over the years. The worldwide
dengue cases reported to WHO increased from 0.5 million cases in 2000 to 5.2 million in 2019. Although the fatality rate is low, there is no specific vaccine or treatment for dengue fever [34]. Thus, surveilling dengue outbreaks becomes crucial. Numerous studies have been conducted on dengue surveillance, most of them are retrospective studies, see [13], [18], [21], [22], [24], [25], [28], [35], [41], and [42]. These studies have contributed to our understanding of dengue outbreaks and have laid the groundwork for developing effective surveillance strategies.

As a subtropical region, Hong Kong is vulnerable to the threat of dengue fever. In August 2018, the Center of Health reported 29 local cases, the highest number in many years. To predict annual dengue incidence in Hong Kong, reference [30] developed a Poisson generalized linear mixed model that utilizes monthly total rainfall and average temperature data. They emphasized that reducing mosquito activities is crucial in preventing dengue outbreaks. This conclusion is echoed by the World Health Organization (WHO), which recommends combating mosquito vectors to control or prevent dengue transmission. Reference [11] reviewed various dengue vector surveillance approaches and noted that ovitrap index (Ovitrap) is an efficient tool for estimating dengue transmission risk and is widely used worldwide.

Since 2000, the Food and Environmental Hygiene Department (FEHD) in Hong Kong started setting up Ovitraps (Figure 1(a)) in selected areas (Figure 1(b)). These devices detect the presence of adult Aedes mosquitoes and based on detected numbers an estimate of their population in the chosen areas can be made. The increase in the number of mosquitoes can be used as an early warning signal of impending dengue outbreaks [1]. Area Ovitrap index for Aedes albopictus (AOI) can be calculated as

\[
\text{AOI} = \frac{\text{Number of the Aedes-positive ovitraps}}{\text{Total number of ovitraps retrieved from a particular area}} \times 100\%.
\]

In reference [27], the authors took the average AOI among different locations in Hong Kong to compute a summary statistic called the MOI and adapted FDA and FPCA algorithms to a monitoring scheme. The proposed method can detect the outbreak of dengue in Aug 2018. One limitation of this approach is that it focuses on temporal surveillance but ignores spatial information, so the location of signals is unknown. To address this limitation, the present paper aims to use the AOI data to detect outbreaks in different areas across Hong Kong by incorporating spatial information into our monitoring scheme.

The online AOI data collected from 2008 to 2018 is available on the website of FEHD. However, as the FEHD expands the coverage of ovitraps every few years, some areas may have fewer historic AOI records than others, making it challenging to monitor newly added locations with limited observations. Reference [6] indicated the spatial correlation of dengue incidence in a medium-sized city in Brazil, which is caused by spatial-dependent dengue transmission factors. In Hong Kong, the temperature, precipitation, population, and other dengue transmission factors in different regions are similar. Hence, the assumption that the AOI data from different areas are correlated is reasonable. If we want to monitor dengue outbreaks in a newly added location with limited historical observations, incorporating historical records from previous regions might be helpful.

Transfer learning is an efficient framework for extracting acquired knowledge from source domains to solve the related tasks in target domains. In data mining and machine learning fields, a domain refers to the combination of the feature space and a marginal probability distribution, while a task consists of a label space and an objective predictive function. The domains and tasks in our problems are a bit different and will be redefined in Section II. Transfer learning has been well-studied in the fields of classification, regression, and clustering, and has been shown to be effective in various applications. For an overview, we refer readers to the surveys conducted by [7] and [19]. In addition to these surveys, [23] provided a comprehensive review of statistical transfer learning that focuses on adopted statistical models and methodologies. They discussed and investigated the utilization of statistical transfer learning in statistical process monitoring (SPM) and quality control.

Transfer learning has been a topic of recent interest for process modeling and monitoring. Several studies have explored the use of transfer learning in this field, including [31], [33], [37], [38], [43], and [44]. Reference [33] adapted a self-taught clustering method to a self-starting control scheme. Reference [31] proposed a transfer learning framework to extract profile-to-profile inter-relationship to improve the monitoring performance. Reference [37] used an instance-based transfer learning approach to aid the distribution inference in the target task with limited data. Reference [38] used Bayesian formulation to facilitate cross-product data sharing. Reference [43] combined probability principal component analysis with transfer learning approach to monitoring processes with data from multiple distributions. Reference [44] proposed a transfer learning-based framework for online unsupervised condition monitoring and anomaly detection. For real applications, reference [32] integrated a parameter-based transfer learning approach with the ordered LASSO to transfer information from old sensors with sufficient historical data to new ones with limited data. Then a generalized likelihood ratio control chart is proposed for monitoring the newly deployed sensors. Reference [36] transferred knowledge from one disease to improve the prediction of another related illness, such as Dengue and Zika, illustrating the application of transfer learning in health surveillance. The spatial correlation of dengue transmission also provides a prerequisite for transfer learning. To our knowledge, there is no research on using transfer learning in prospective dengue surveillance, leaving a research gap.
Our objective in this paper is to build a monitoring scheme to detect dengue outbreaks in target areas with limited observations in Hong Kong. Limited data restricts the application of traditional parametric control charts, and waiting for more data will take months during which we prefer to monitor for outbreaks. To address this problem, we propose a novel transfer learning-based method. First, we introduce the data structure which includes target and source domains in Section II-A. Next, we describe the fundamental model for our method and present the transfer learning algorithm with the corresponding estimator. Finally, we propose a monitoring scheme based on the transfer learning estimator. We will also answer the three main issues in transfer learning in the following subsections: what to transfer, how to transfer, and when to transfer.

II. METHODOLOGY

This section defines the problem and introduces our transfer-learning-based method. First, we introduce the data structure which includes target and source domains in Section II-A. Next, we describe the fundamental model for our method and present the transfer learning algorithm with the corresponding estimator. Finally, we propose a monitoring scheme based on the transfer learning estimator. We will also answer the three main issues in transfer learning in the following subsections: what to transfer, how to transfer, and when to transfer.

A. SOURCE & TARGET DOMAINS

The following introduction regards the source and target domains as the AOI data sets collected from old and newly added locations. The source and target domains can be called source and target processes and source and target data in the following sections. We aim to use multiple source domains to solve target tasks in one target domain. We assume the underlying distributions of all domains are unknown and that enough source data are available for accurately estimating their underlying distributions. But the target domain only has limited observations. The target and source data are shown in Equation 1.

\[
\begin{align*}
X_0 &= (X_{0,1}, X_{0,2}, \ldots, X_{0,n_0}); \\
X_1 &= (X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}); \\
& \vdots \\
X_K &= (X_{K,1}, X_{K,2}, \ldots, X_{K,n_K}),
\end{align*}
\]

where \(X_0\) is the target domain and \(X_k, (k = 1, \ldots, K)\) are the \(K\) source domains. \(n_0\) is the number of target observations, which is much smaller than \(n_k (k = 1, \ldots, K)\), and too small to estimate parameters for the target domain accurately. In the dengue case, \(X_0\) is the AOI data collected from a newly added location, and the other \(K\) data sets are AOI records from the old locations.

We assume that the observations \(X_{k,i} (i = 1, \ldots, n_k)\) in the \(k\)-th \((k = 0, \ldots, K)\) domain follow a multivariate normal distribution with \(p\) variables

\[
X_{k,i} \sim N(\mu_k, \Sigma_p),
\]

where \(\mu_k\) is the mean vector in the \(k\)-th distribution, there are \(K + 1\) distributions with the same dimension \(p\) but different expectations. According to the FDA results in [27], a seasonal period in AOI data is one year with 12 data points for each month. We can treat them as \(p = 12\) variables, which are constant across all areas. The expected AOI values \((\mu_k, k = 0, \ldots, K)\) in each location are different. The covariance matrix \(\Sigma_p\) is assumed to be the same for all distributions. In other words, variables in different domains...
share the same dependence structure. This makes sense in the dengue case as the relationship among variables is related to the autocorrelation in the AOI series. In other words, the series from different locations share similar seasonality and trend.

$K + 1$ expectations from different distributions can be combined into a $(K + 1) \times p$ matrix:

$$U = \begin{bmatrix}
\mu_0 \\
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_K
\end{bmatrix} = \begin{bmatrix}
\mu_{0,1} & \mu_{0,2} & \cdots & \mu_{0,p} \\
\mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,p} \\
\mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,p}
\end{bmatrix} \tag{3}$$

Each row of $U$ represents a $p$-dimensional location parameter from Equation 2. In the AOI data, it is an expectation vector with 12 elements for each area. Each column represents the expectation for variable $j$ across all processes. For example, column 1 shows the expected AOI levels in January across all $K + 1$ locations. Given that dengue factors in different areas show spatial correlation, it is reasonable to assume that the AOI data from different areas are correlated. The spatial correlation confirms that the information from source domains is worth transferring to the target domain, which answers the question “when to transfer”.

The $K + 1$ rows in Equation 3 are interdependent. Let $\mu_{\cdot, j} = (\mu_{0,j}, \mu_{1,j}, \ldots, \mu_{K,j})'$ denotes the $j$-th column of $U$. The $K + 1$ components in $\mu_{\cdot, j}$ are correlated. The AOI values from different locations have similar levels in the same month, so the $K + 1$ components in the same column $\mu_{\cdot, j}$ share the same expectation. That is $E[\mu_{0,1}] = E[\mu_{1,1}], \ldots, = E[\mu_{K,1}]$. To connect all $K + 1$ domains, we use multivariate Gaussian priors for each variable $j$

$$\mu_{\cdot, j} \sim N(\mu^j, \Sigma_{K+1}), j = 1, \ldots, p, \tag{4}$$

where $\mu^j$ is the mean vector with identical elements $\mu$, which is unknown and needs to be estimated because of the similar seasonality and trend of AOI data in different locations, the identical assumption is reasonable. For each column in Equation 3, $E[\mu_{0,1}] = E[\mu_{1,1}], \ldots, = E[\mu_{K,1}] = \mu^j$. Then each row in Equation 3 has the same expectation, $E[\mu_0] = E[\mu_1], \ldots, = E[\mu_K] = (\mu^1, \mu^2, \ldots, \mu^p)'$. The expectations of AOI data from different locations are the same. The covariance matrix $\Sigma_{K+1}$ in prior distributions represents the hidden dependency structure among $K + 1$ rows in Equation 3. In the dengue case, $\Sigma_{K+1}$ is the spatial correlation among different locations, which is constant in $p$ prior distributions. After building the connection via Gaussian priors, we can effectively transfer useful source domain information to the target domain.

Our objective is to detect changes in AOI levels for newly added locations. This consists of two tasks, modeling the underlying distribution and monitoring online processes. Parameter estimation and control chart design based on limited target observations are major challenges. With correlated source domains, we can improve the estimation results for the target process using the information from source domains. We can better understand the in-control condition of the target process and design a reliable control chart based on it. The answers for “what to transfer” are the parameters and monitoring schemes.

### B. TRANSFER LEARNING ALGORITHM AND ESTIMATOR

This section explains the transfer learning algorithm in detail to answer the “how to transfer” problem. As we assumed in the model, the underlying distribution for each domain is unknown and needs to be estimated. Generally, the sample mean computed by $\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{i,k}$ is used to estimate the location parameter for all $K + 1$ domains. These estimated expectations can be aggregated into a matrix as follow, which is the estimate of Equation 3.

$$\hat{U} = \begin{bmatrix}
\hat{\mu}_0 \\
\hat{\mu}_1 \\
\hat{\mu}_2 \\
\vdots \\
\hat{\mu}_K
\end{bmatrix} = \begin{bmatrix}
\hat{\mu}_{0,1} & \hat{\mu}_{0,2} & \cdots & \hat{\mu}_{0,p} \\
\hat{\mu}_{1,1} & \hat{\mu}_{1,2} & \cdots & \hat{\mu}_{1,p} \\
\hat{\mu}_{2,1} & \hat{\mu}_{2,2} & \cdots & \hat{\mu}_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\mu}_{K,1} & \hat{\mu}_{K,2} & \cdots & \hat{\mu}_{K,p}
\end{bmatrix} \tag{5}$$

As we assume enough source data, $\hat{\mu}_k, k = 1, \ldots, K$ are accurate estimates for source processes. But $n_0$ is a small positive integer, so $\hat{\mu}_0$ is an inaccurate estimate for the target data. The covariance matrix $\Sigma_{K+1}$ in the prior distributions in Equation 4 connects the expectations of different domains and reflects the dependency structure among these domains. So we select the most highly correlated source domain, then learn and transfer useful information from the selected data set to improve the estimation accuracy for the target process.

To estimate the underlying correlation among different domains, we first estimated the location parameters of $N(\mu^j, \Sigma_{K+1}), j = 1, \ldots, p$. As discussed in Section II-A, each column in Equation 3 has the same expectation, so that $\mu^j$ can be estimated by

$$\hat{\mu}^j = \frac{\sum_{k=0}^{K} \hat{\mu}_{k,j}}{K + 1}, \hat{\mu}^j = (\hat{\mu}_1^j, \hat{\mu}_2^j, \ldots, \hat{\mu}_p^j)' , j = 1, \ldots, p. \tag{6}$$

Then the scatter matrix for $\Sigma_{K+1}$ is computed as

$$\hat{\Sigma}_{K+1} = \frac{1}{p} \sum_{j=1}^{p} (\hat{\mu}_{\cdot, j} - \hat{\mu}^j)(\hat{\mu}_{\cdot, j} - \hat{\mu}^j)' \tag{7}$$

When $p < K + 1$, the scatter matrix is an inaccurate estimator for $\Sigma_{K+1}$, but it can still reflect the correlation among different domains. So we use the scatter matrix as a reference to determine which source domain should be selected. Elements in the first row (or the first column) of $\Sigma_{K+1}$ show the dependency between the target domain and other $K$ source domains. We can select the source data ($X_s$) with the most significant positive correlation to learn and transfer its sample mean $\hat{\mu}_s$.

The transfer learning-based estimator for $\mu_0$ is

$$\hat{\mu}_0 = w \hat{\mu}_0 + (1 - w) \hat{\mu}_s \tag{8}$$
where \( \hat{\mu}_0 \) is the sample mean vector of \( X_0 \), and \( \hat{\mu}_s \) is the sample mean of the selected source domain, which is highly correlated with the target process. \( w \in [0, 1] \) is the weight parameter. If \( w < 1 \), \( E[\hat{\mu}_0] \neq \mu_0 \), and \( \hat{\mu}_0 \) is a biased estimator with the bias \((1 - w)(\mu_s - \mu_0)\). We derive the expectation of mean square error (MSE) for \( \hat{\mu}_0 \) and \( \mu_0 \) as

\[
E[MSE] = \frac{1}{p} \sum_{j=1}^{p} \sigma^2_{0,j},
\]

\[
E[MSE] = \frac{1}{p} \sum_{j=1}^{p} \left[ w^2 \sigma^2_{0,j} \frac{n_0}{n_s} + (1 - w)^2 \sigma^2_{s,j} \frac{n_s}{n_s} 
+ 2w(1 - w) \rho_{0,s,0,j} \rho_{s,j} \frac{\sqrt{n_0n_s}}{\sqrt{n_0n_s}} 
+ (1 - w)^2 (\mu_{s,j} - \mu_{0,j})^2 \right],
\]

where \( \sigma^2_{0,j} \) and \( \sigma^2_{s,j} \) are the variance for variable \( j \) in the target domain and selected processes respectively, and \( \rho_{0,s} \) is the correlation. Derivation of Equation 9 are shown in the Appendix. According to the definition of the prior distribution, \( E[\mu_0 - \mu_s] = 0 \). Therefore, when \( n_0 \) is small, selecting a proper source domain with strong correlation \( \rho \) and choosing a proper \( w \) value can ensure \( E[MSE - MSE] \geq 0 \). Confirming that the transferred estimator is more accurate than the sample mean of the source domain. Solving this inequality (see the Appendix for details), we obtain a bound on \( w \) such that \( E[MSE - MSE] \geq 0 \):

\[
w \in \left[ \frac{2(1 - \rho_{0,s})n_0n_s + n_0 - n_s}{n_0 + n_s + 2(1 - \rho_{0,s})n_0n_s - 2\rho_{0,s}\sqrt{n_0n_s}}, 1 \right].
\]

Figure 2 plots the bound on \( w \) for various values of \( n_0 \) and \( \rho \). Stronger correlation between the source and target domain allows for a larger weight on \( \hat{\mu}_s \) (i.e., a smaller \( w \)). For a very strong correlation and small sample size, it is even possible to only use the source domain data for estimation. As \( n_0 \) increases, the range for selecting \( w \) becomes smaller and converges to 0. Note that this plot only provides a range for selecting \( w \) instead of the optimal choice. We can compute the optimal \( w \) value by setting the partial derivative of \( E[MSE] \) with respect to \( w \) equal to 0. These derivations can be found in Appendix. The optimal value of \( w \) is:

\[
w = \frac{n_0 + 2(1 - \rho_{0,s})n_0n_s - \rho_{0,s}\sqrt{n_0n_s}}{n_0 + n_s + 2(1 - \rho_{0,s})n_0n_s - 2\rho_{0,s}\sqrt{n_0n_s}}.
\]

Algorithm 1 summarizes the detailed selection steps.

**Algorithm 1 Parameter Estimation via Transfer Learning**

**Input** \( X_k (k = 0, \ldots, K), w \)

**Output** \( \hat{\mu}_0 \)

**Step 1.** Computing the sample mean for each domain and aggregated them as a \((K + 1) \times p\) matrix by Equation 5;

**Step 2.** Computing the sample mean for each prior distribution using Equation 6;

**Step 3.** Estimating the covariance matrix \( \hat{\Sigma}_{K+1} \) in the prior distribution by Equation 7;

**Step 4.** In the first row of \( \hat{\Sigma}_{K+1}, \hat{\sigma}_1.: = (\hat{\sigma}_{1,1}, \hat{\sigma}_{1,2}, \ldots, \hat{\sigma}_{1,K+1}) \), take \( s = \arg \max_{q=1}^{K+1} \hat{\sigma}_{1,q} \) where \( \hat{\sigma}_{1,K+1} = 2, 3, \ldots, K + 1 \) as the selected source data \( X_s \);

**Step 5.** Use \( w, \hat{\mu}_0 \) and \( \hat{\mu}_s \) to compute \( \hat{\mu}_0 \) by Equation 8.

### C. PROCESS MONITORING BASED ON TRANSFER LEARNING

Using the Algorithm 1, we get a more accurate estimate for the target mean vector and solve the modeling tasks. Another target task is building a reliable monitoring scheme to detect changes in the target process. We modify the Hotelling \( T^2 \) control chart to incorporate the transfer learning-based parameters. The classical \( T^2 \) chart \( (T^2) \) for an individual observation \( X_{k,i} \) in process \( k \) is

\[
T^2_{k,i} = (X_{k,i} - \hat{\mu}_k)^T \hat{\Sigma}_p^{-1} (X_{k,i} - \hat{\mu}_k),
\]

where \( \hat{\mu}_k \) is the sample mean for \( k \)-th process.

In Equation 2, we assume that all domains share the same covariance matrix, it can be estimated based on all \( K + 1 \) data sets as \( \hat{\Sigma}_p = (\sum_{k=0}^{K} \sum_{i=1}^{n_k} (X_{k,i} - \hat{\mu}_k)(X_{k,i} - \hat{\mu}_k)^T)/(\sum_{k=0}^{K} n_k - (K + 1)) \). It is accurate with a large amount of pooled data.

To build control charts based on the estimator in Equation 8, we replace the sample mean and \( \hat{\Sigma}_p \) in Equation 12 with \( \hat{\mu}_k \) and \( \hat{\Sigma}_p \) respectively. Where \( \hat{\mu}_k \) is the transfer learning-based mean estimate for each domain based on Algorithm 1, and \( \hat{\Sigma}_p = (\sum_{k=0}^{K} \sum_{i=1}^{n_k} (X_{k,i} - \hat{\mu}_k)(X_{k,i} - \hat{\mu}_k)^T)/(\sum_{k=0}^{K} n_k - (K + 1)) \) is the covariance matrix corresponding to the transfer learning-based estimator based on pooled data. The transfer learning-based \( T^2 \) statistics is computed as

\[
\hat{T}^2_{k,i} = (X_{k,i} - \hat{\mu}_k)^T \hat{\Sigma}_p^{-1} (X_{k,i} - \hat{\mu}_k).
\]

We can compute the \( \hat{T}^2_{k,i} \) statistics for all source domains, and it is large enough to build an experimental distribution \( \hat{F}_Y(y) \) for \( \hat{T}^2_{k,i} \). We can use the quantile of \( \hat{F}_Y(y) \) to estimate the control limits as

\[
\hat{CL} = \hat{F}_Y^{-1}(\alpha),
\]

where \( \alpha \) is the false alarm rate (FAR) which makes in-control average run length \( ARL_0 = \frac{1}{\alpha} \) theoretically. The transfer learning-based control chart is named the \( TL - T^2 \) chart.
The second target task of building an efficient monitoring scheme is solved now. For the classical $T^2$ chart, we can make an empirical distribution $F_Y(y)$ based on the source data sets and estimate the control limits based on its quantile $CL = F_Y^{-1}(w)$.

### III. SIMULATION STUDIES

In this section, we employ simulations to investigate the performance of the proposed method in both offline estimation and online monitoring. We vary the parameters in different scenarios, so the results also provide a guideline for designing the control charts. To measure the estimation accuracy, we compare the $MSE$ of $\hat{\mu}_k$ with the $MSE$ of $\hat{\mu}_k$. We use the $ARL$ to evaluate the in-control and out-of-control performance of the $TL - T^2$ chart.

We set the number of source domains as $K = 10, 20$ with $p = 10, 20, 30$ variables. The mean values $\mu_j, j = 1, \ldots, p$ in the $p$ prior distributions are randomly drawn from a uniform distribution $U(0, 4)$ in each simulation run. The prior covariance matrix is fixed as $\Sigma_{K+1} = (\sigma_{l,m}|1 \leq l, m \leq (K+1))$, where $\sigma_{l,m} = 0.9^{l-m}$. Then mean vectors $\hat{\mu}_k$ for each domain are generated from prior distributions $N(\mu_j, \Sigma_{K+1}), j = 1, \ldots, p$. The constant covariance matrix in Equation 2 is set as $\Sigma_p = (\sigma_{l,m}|1 \leq l, m \leq p)$, where $\sigma_{l,m} = 0.7^{l-m}$. Data sets for each domain are drawn from the corresponding multivariate normal distributions $N(\mu_k, \Sigma_p)$. The numerical results in this section are all based on 3000 simulations.

#### A. ESTIMATION PERFORMANCE

As we discussed in Section II-B and Equation (11), weight $w$ is an important factor for the accuracy of $\hat{\mu}_0$. Hence, we explore the effect of $w$ on the estimation accuracy and provide guidelines for selecting proper $w$ values. We use $n_0 = 5, 10$ data points for the target domain. We expected that the transfer learning-based estimator could achieve smaller estimation error with proper weights. The $MSE$ for $\hat{\mu}_k$ with various values of $w$ are shown in Figure 3.

Figure 3 plots the $w$ on the x-axis and the $\tilde{MSE}$ on the y-axis as well as the optimal value for $w$ as obtained from Eq. (11). Each line represents one combination of $K$ and $p$. $MSE$ is the horizontal line in each plot as a reference. When $MSE$ is below the horizontal $MSE$ line, it is beneficial to transfer knowledge. All curves share the same U-shape trend, $MSE$ values first decrease and then increase. When $w = 0.1$, the estimator is affected by the bias significantly and has a large estimation error. When $w = 1$, the transfer learning-based estimator equals the sample mean, $\tilde{MSE} = MSE$. When $n_0 = 5$, see Figure 3a, $\tilde{\mu}_0$ can achieve the minimal estimation error with $w = 0.501$. As shown in Figure 3b, when $n_0 = 10$, the lowest values of $MSE$ are achieved at $w = 0.678$. More target data can reduce the minimal estimation error. The estimation accuracy is significantly improved by learning useful knowledge from the selected source domain with a properly selected weight when the target data are limited. These observations also are in line with the result in Figure 2 that when $\rho = 0.9, n_0 = 5$, the range of $w$ is from 1 to 0. When $n_0 = 10$, $w$ should be selected between 0.4 and 1.

With different combinations of $K$ and $p$, the minimal values of $MSE$ curves and the corresponding weight are slightly different. When $K < p$, the $MSE$ curves are similar and overlapped. And the proposed estimator can achieve the minimal estimation error around 0.1 with $w = 0.5$ in Figure 3a. Because when $K < p$, the result from Equation 7 is closer to the real $\Sigma_{K+1}$, and the most relevant source data is more likely to be selected to improve the estimation accuracy. When $K \geq p$, $\tilde{\mu}_0$ shows a more significant estimation error, but it is still more accurate than $\hat{\mu}_0$ with proper weight. This difference is caused by the inaccurate estimate of $\Sigma_{K+1}$ and its negative effect on domain selection decisions.

The difference between Figure 3a and Figure 3b highlights the impact of $n_0$ on the accuracy of $\hat{\mu}_0$. With more observations, $\tilde{\mu}_0$ becomes more accurate. The minimal $MSE$ achieved for $n_0 = 10$ is smaller than that achieved for $n_0 = 5$. Additionally, the corresponding weight is different, with $w = 0.5$ for $n_0 = 5$ and $w = 0.7$ for $n_0 = 10$. This result suggests that determining the optimal weight should depend on the number of target data ($n_0$).

$\hat{\mu}_0$ represents a trade-off between estimation error and bias. With few target data, $\hat{\mu}_0$ exhibits a significant estimation error, so we assign small weights to reduce the estimation error while allowing the bias to cancel out some of the estimation errors. With more target data, the sample mean has a smaller estimation error, so we assign more weight to it to reduce the bias. Overall, we recommend that practitioners determine the weight based on the number of target data. When $n_0$ is very small, we recommend using small weights around 0.5. When more target data are available, we suggest using larger values of $w$. We recommend practitioners use the specific $n_0$ and expected $\rho$ to simulate the optimal $w$ in different applications.

Equation 9 and Figure 3 demonstrate that $\tilde{MSE}$ and $\tilde{MSE}$ are influenced by $n_0$. As $n_0$ becomes larger, the classical estimator becomes more accurate than the transfer learning-based estimator, and $\tilde{\mu}_0$ is not recommended for larger $n_0$. Therefore, we compare $MSE$ and $\tilde{MSE}$ with various $n_0$ to investigate the effect of sample size. We use the same settings for $K$ and $p$ as in Figure 3, and choose $w = 0.5$ and 0.7 to include the effect of weight in the comparison. The expected results are that both $\tilde{MSE}$ and $\tilde{MSE}$ would decrease with increasing $n_0$, and that $MSE$ values would be greater than $\tilde{MSE}$ with small $n_0$. Figure 4 plots the values of $MSE$, $\tilde{MSE}$ with $w = 0.5$, and $\tilde{MSE}$ with $w = 0.7$ on the y-axis as the number of observations ($n_0$) increases.

All six plots demonstrate that $MSE$ and $\tilde{MSE}$ decrease as the number of target data points increases and converge to a small value. This suggests that both estimators can achieve more accurate estimates with more observations. When the number of target data points is limited (i.e., $n_0 = 5$), both estimators have larger estimation errors, but $\tilde{\mu}_0$ is more accurate than $\hat{\mu}_0$. As $n_0$ increases, $\tilde{\mu}_0$ outperforms $\hat{\mu}_0$ significantly. The minimum $MSE$ is around 0.05 with $w = 0.5$, which is
caused by the bias term, \((1 - w)(\mu_x - \mu_0)\). The minimum \(\overline{MSE}\) is smaller with \(w = 0.7\), indicating that using a larger value of \(w\) can significantly reduce the bias.

When \(n_0 \leq 14\), \(\hat{\mu}_0\) has a smaller estimation error with both \(w = 0.5\) and \(w = 0.7\) in most scenarios. However, when there are more than 14 observations, \(\hat{\mu}_0\) with \(w = 0.5\) becomes inaccurate. Meanwhile, \(\hat{\mu}_0\) with \(w = 0.7\) has the lowest estimation error when \(8 < n_0 \leq 23\). The sample mean is the best estimator when more than 25 data points are available. While \(K\) and \(p\) have an insignificant effect on the estimation error, the transfer learning estimator loses effectiveness faster when \(K > p\).

Equation 9 shows that \(\overline{MSE}\) depends on the correlation between the target domain and the selected domain, which is represented by the term \(2w(1 - w)(\rho_{0, j}\sigma_{0, j}/\sigma_{0, j}/\sqrt{\mu_0\sigma_{0, j}})\). The term \(2w(1 - w)\) has its maximum value of 0.5 at \(w = 0.5\). In this simulation study, we assume that all variables have a variance of one, and \(n_0\) is a large value. Therefore, \((\rho_{0, j}\sigma_{0, j}/\sigma_{0, j}/\sqrt{\mu_0\sigma_{0, j}})\) is still small when \(\rho\) is large. As the effect of \(\rho\) on \(w\) has been demonstrated in Figure 2, we do not conduct any additional simulations to investigate their relationship.

Overall, the proposed estimator \(\hat{\mu}_0\) can produce a more accurate estimate with limited target data, particularly in higher-dimensional scenarios. Larger weights are recommended to improve estimation accuracy. Simulation is recommended to determine the optimal \(w\) based on a specific \(n_0\) and \(\rho\).

**B. PERFORMANCE OF CONTROL CHARTS**

In this section, we compare the in-control and out-of-control performance of the \(TL - T^2\) control chart with the Hotelling \(T^2\) control chart as defined in Equation 12. The Hotelling \(T^2\) control chart is chosen as the comparator for two reasons: first, it is a well-known traditional method that is easy to understand and interpret, and second, the proposed method extends the Hotelling \(T^2\) method with the transfer learning estimate as a major difference. We consider scenarios where \(K = 10\) and \(p = 10, 20, 30\). We set \(w = 0.5\) when the target domain has \(n_0 = 5\) observations and \(w = 0.7\) when \(n_0 = 10\) (as determined in Section III-A). The control limits for both charts are computed using the corresponding empirical distributions, as discussed in Section II-C. We expect the \(TL - T^2\) chart to have an \(ARL_0\) closer to theoretical values and a smaller detection delay under various scenarios. Because it is based on more accurate estimators.

Figure 5 shows the difference between the simulated \(ARL_0\) and the theoretical \(ARL_0 = 1/\alpha\) on the \(y\)-axis, and \(\alpha\) from 0.01 to 0.1 on the \(x\)-axis. The curves closer to zero indicate that the simulation results are closer to the design values. The hollow dots represent the \(ARL_0\) deviation of the \(T^2\) chart, and the asterisks represent the \(ARL_0\) deviation of the \(TL - T^2\) chart. As \(\alpha\) increases, both methods converge to the theoretical values, but the \(TL - T^2\) chart converges faster. The negative difference in all scenarios indicates that both charts have smaller \(ARL_0\) than expected values. The deviations for both charts are more prominent in higher-dimensional scenarios with small \(\alpha\). Since \(n_0\) is small, the sample mean is an inaccurate estimator for the target data, resulting in a more significant deviation in the \(ARL_0\) of the \(T^2\) chart. The \(TL - T^2\) chart can achieve a smaller deviation in all scenarios because of the more accurate estimate.

The difference between Figure 5a and 5b suggests that using larger weights with more target data can reduce the deviation of \(ARL_0\), resulting in faster convergence. This could be because the transfer learning-based estimator with smaller weights has a larger bias. Additionally, more variables can slow down the convergence, and the degeneration is more significant with small \(w\) and small \(n_0\). Overall, the in-control \(ARL\) values of the proposed \(TL - T^2\) chart are closer to \(\frac{1}{\alpha}\) than those of the \(T^2\) chart, and they can converge to the expected value quickly. If practitioners are sensitive to false alarms, our proposed method is a better alternative to the \(T^2\) chart. Practitioners can update control limits by varying \(\alpha\) in Equation 14 to achieve a preferred target \(ARL_0\).

We also compare the out-of-control performance of the \(TL - T^2\) control chart with the \(T^2\) chart by monitoring changes in the mean vector \(\hat{\mu}_0\), where \(\hat{\mu}_0 = \mu_0 + \delta\) and \(\delta\) is the shift size. The sets of \(K, p, w,\) and \(n_0\) are the same as in Figure 5. We set \(\alpha = 0.02\) for the \(TL - T^2\) chart since we assume that the target domain has limited observations, so we
do not consider very large $ARL_0$. To make the $ARL_0$ values comparable, we adjust the $\alpha$ for the $T^2$ chart. The results for both methods are presented in Table 1.

With $n_0 = 5$ target observations, the $TL - T^2$ chart can detect medium and large changes more efficiently than the $T^2$ chart. The $T^2$ chart can detect changes with size $\delta = 0.5$ faster than our proposed method. One reasonable explanation is that the $T^2$ chart has smaller $ARL_0$ than the $TL - T^2$ chart. As $ARL_0$ is a random variable, it is impossible to make $ARL_0$ for two methods precisely the same, so we keep them similar. Both charts show significant detection delay for such minor changes. Both methods’ $ARL_1$ decreases with increasing shift sizes. When $\delta \geq 1$, our proposed method outperforms the $T^2$ chart with a smaller detection delay. When $\delta = 3$, both charts can signal a shift as soon as it occurs. With more target observations, $n_0 = 10$, both methods can achieve larger $ARL_0$, and the $TL - T^2$ chart is more sensitive to medium and large shifts. The $TL - T^2$ method has a smaller $ARL_1$ with $p = 10$ and $p = 20$. Additionally, it has a smaller $ARL_0$ in both scenarios, which confirms our explanation that methods with smaller $ARL_0$ have less detection delay for small shifts. Both charts perform better in higher-dimensional scenarios because shifts in more variables can make the cumulative changes in $T^2$ type statistics more apparent.

Overall, our proposed method can achieve more accurate $ARL_0$ values, especially with larger $w$ in higher-dimensional scenarios. Table 2 summarizes the strengths and weaknesses of our proposed chart. Our method is preferable to the $T^2$ chart when we have limited target domain data and extensive source domain data that is (believed) to be correlated with the source domain. We recommend practitioners determine and adjust control limits based on their preferred
in-control performances. The $TL - T^2$ chart can efficiently detect changes in the target process. When the target data is limited, and the practitioner is sensitive to false alarms, we recommend using the $TL - T^2$ chart with a large weight $w$ to monitor for shifts in the target process.

**IV. CASE STUDY**

This section applies the proposed method to monitor the AOI data from a new area in Hong Kong. Hong Kong experienced a dengue outbreak in 2018, so we expect the proposed method to detect a signal in that year. Since AOI data are collected monthly, one location has 12 records in one year. They can be treated as time series, but the temporal correlation is challenging to model. Another alternative is treating the one-year AOI data as an observation with $p = 12$ variables, and the temporal correlation is the dependence among variables, which can be estimated by $\hat{\Sigma}_p$. The target domain only has eight observations from 2011 to 2018. For the $K = 25$ source domains, each includes 11 records from 2008 to 2018. We assume all source data and the first $n_0 = 5$ target observations are in control.

First, we use the Algorithm 1 to estimate the transfer learning-based mean vector for the target area. Figure 6 shows the in-control target data, the sample mean vector, and the estimated mean vector. The solid black lines are the annual AOI data; the x-axis indicates 12 variables. The sample mean is the red dotted line, and the green line shows the transfer learned mean vector. The temporal correlation and seasonality are obvious, with smaller AOI values in winter and the highest AOI values in summer. Both estimators can reflect the seasonal pattern of AOI data.
We compute 275 $\tilde{T}^2$ statistics based on 25 source areas and take the quantile at $FAR = 0.1$ as the control limit for the $T^2$ and $TL - T^2$ charts. Since one monitoring statistic represents the condition for one year, $ARL_0 = 10$ is a reasonable and practicable choice for this case. Figure 7 shows the classical $T^2$ and the $TL - T^2$ charts. Both charts can detect a signal in 2018, consistent with the reported outbreak. The $T^2$ chart shows a monitoring statistic in 2017 close to the control limits, but there was no outbreak that year. The signal in 2017 will be treated as a false alarm. This is in line with the results in Table 1 which showed that the $T^2$ chart based on limited target data has shorter than expected $ARL_0$ values.

The $TL - T^2$ chart requires a one-year AOI vector to compute the monitoring statistic, which causes detection delay. The transfer learning-based estimator can model and remove the underlying seasonality of AOI data. Complementary to the $TL - T^2$ control chart, we applied a univariate Shewhart control chart to monitor the residuals. Since AOI values are near zero in winter, indicating a low risk of outbreaks, we removed the records in January, February, March, November, and December. The rest residuals show a heavy tail, either standardized by the sample mean or the transfer learning-based estimator, see Figure 8. So we use the sample interquartile range as a robust standard deviation estimator. The standard deviations for the residuals are $\hat{\sigma} = 0.046$, and $\tilde{\sigma} = 0.052$ for the two different estimators respectively. In this case, the expected average time to signal is set at 70 months (10 years without winter), so $\alpha = 1/70$, we are interested in increasing changes, so the control limits are 0.115 and 0.123. Each Shewart control chart in Figure 9.

### TABLE 2. Pros and cons of $TL - T^2$ control charts.

<table>
<thead>
<tr>
<th>Pros</th>
<th>$TL - T^2$</th>
<th>$T^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>More accurate parameter estimation with the transfer learning algorithm.</td>
<td>Easy to understand and apply.</td>
</tr>
<tr>
<td></td>
<td>Efficient for monitoring processes with limited historical data.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robust in-control performance.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usage of related knowledge.</td>
<td></td>
</tr>
<tr>
<td>Cons</td>
<td>Requires strong correlation among source and target data sets.</td>
<td>Requires a large amount of in-control data.</td>
</tr>
<tr>
<td></td>
<td>Performance depends on the transferred knowledge.</td>
<td>Based on multivariate normal distributions.</td>
</tr>
<tr>
<td></td>
<td>Based on multivariate normal distributions.</td>
<td>Less efficient with limited historical data</td>
</tr>
</tbody>
</table>

(a) Hotelling $T^2$ control chart  
(b) $TL - T^2$ control chart

**FIGURE 7.** Control charts.

**FIGURE 8.** The Q-Q plots for residuals.
shows three signals, one in May 2017, one in July 2017, and one in April 2018. However, there was only one outbreak in 2018. The signals in 2017 are false alarms. But this result is consistent with the multivariate control charts in Figure 7, that both 2017 and 2018 have large $T^2$ statistics.

This case study suggests that the proposed methods can be used to learn helpful information from a similar process, make more accurate estimators, and monitor the target process with limited observations. It is applicable in the health surveillance field.

V. CONCLUSION

This paper proposes a transfer learning-based strategy for dengue surveillance in multiple areas where the available historical observations are limited for some areas. We proposed a transfer learning algorithm to connect the dengue data from areas with a lot of historical data (source domain) to locations with limited historical data (target domain). The knowledge from source domain is used to improve the estimation and monitoring scheme for the target domain. The developed auxiliary estimator can provide accurate estimates even with very limited target data, using both target and source domain data. Using this transfer-learned parameter estimate, the corresponding control chart is sensitive to shifts in mean vectors. As far as we know, this is the first research paper that uses transfer learning for prospective dengue surveillance; both the transfer learning algorithm and the auxiliary estimator are innovative solutions for dengue surveillance and can be easily extended to other diseases and fields. The proposed method is recommended for monitoring high-dimensional processes where we have extended historical data for some processes and only limited historical data available ($n_0 \leq 25$) for the target process. Our proposed method can learn and build control charts using knowledge from related processes.

Comparisons of $MSE$ show that our proposed estimator is more accurate than the sample mean vector with limited target observations. And the optimal choice of weights depends on the number of target observations and correlations. Larger weights are recommended since they can achieve minor estimation errors under various scenarios. The proposed estimator loses its advantage when the number of target data exceeds 25. Based on a more accurate estimate, the $TL - T^2$ has better in-control performance than the Hotelling $T^2$ chart. The $ARL_1$ results show that the proposed method can efficiently detect mean shifts. A dengue surveillance example illustrates the applicability of the proposed method in monitoring a real process with limited observations. Conclusively, we recommend using the $TL - T^2$ control chart to monitor a process with limited historical data.

The limitation of the proposed methods is evident. The proposed transfer learning-based estimator only learns from one source domain, which may ignore helpful information from other sources. Therefore, one potential future research direction is updating the selection algorithm to include more valuable knowledge. And we use a Gaussian prior distribution to connect different processes, which may not be accurate in practice. Finding a more general method to combine other domains is worth considering. Adapting the transfer learning algorithm to run rules $T^2$ chart, CUSUM chart, and EWMA control charts for monitoring small shifts is an interesting extension of the proposed method.

APPENDIX

The expectation of mean square error (MSE) for $\hat{\mu}_{0}$ is derived as:

$$E[MSE] = E\left[\frac{1}{p} \sum_{j=1}^{p} (\hat{\mu}_{0,j} - \mu_{0,j})^2\right]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \left[\text{Var}(\hat{\mu}_{0,j} - \mu_{0,j}) + E[(\hat{\mu}_{0,j} - \mu_{0,j})^2]\right]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \left[\text{Var}\left(\frac{\sum_{i=1}^{n_0} x_{0,i,j}}{n_0}\right)\right]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \frac{\alpha^2_{0,j}}{n_0}. \ (15)$$

The expectation of mean square error (MSE) for $\hat{\mu}_{0}$ is

$$E[\hat{MSE}] = E\left[\frac{1}{p} \sum_{j=1}^{p} (\hat{\mu}_{0,j} - \mu_{0,j})^2\right]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \left[\text{Var}(\hat{\mu}_{0,j} - \mu_{0,j}) + E[(\hat{\mu}_{0,j} - \mu_{0,j})^2]\right]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \left[\text{Var}(w\hat{\mu}_{0,j} + (1 - w)\hat{\mu}_{s,j} - \mu_{0,j})\right]$$
To study when $E[\overline{MSE}] > E[\overline{MSE}]$ we define $D$ equal to the difference as:

$$D = E[\overline{MSE}] - E[\overline{MSE}]$$

$$= \frac{1}{p} \sum_{j=1}^{p} \frac{\sigma_{0,j}^2}{n_0} - \frac{1}{p} \sum_{j=1}^{p} \left( 2w - n_0 \right) + \left( 1-w \right)^2 \frac{\sigma_{s,j}^2}{n_s}$$

$$+ 2w \left( 1-w \right) \frac{\rho_{0,s} \sigma_{s,j} \sigma_{s,j}}{\sqrt{n_0 n_s}}$$

$$+ (1-w)^2 (\mu_{s,j} - \mu_{0,j})^2$$

$$= \left( 1-w \right) \frac{1}{p} \sum_{j=1}^{p} D_j,$$

where

$$D_j = \frac{1}{n_0} \frac{\sigma_{s,j}^2}{n_s} - \frac{1}{n_0} \frac{(1-w)(\mu_{s,j} - \mu_{0,j})^2}{n_s}$$

$$- 2w \frac{\rho_{0,s} \sigma_{s,j} \sigma_{s,j}}{\sqrt{n_0 n_s}} - (1-w)(\mu_{s,j} - \mu_{0,j})^2.$$

Where $D_j \geq 0$ for all $j$ is a sufficient condition for $D \geq 0$.

So we derive a range for $w$ in order to satisfy $D_j \geq 0$. Without losing of generality, we assume that $\sigma_{0,j} = \sigma_{s,j} = 1$ for $j = 1, \ldots, p$, $D_j \geq 0$; can be simplified as

$$D_j \geq 0 \implies \frac{1}{n_0} \frac{1}{n_s} \frac{(1-w)(\mu_{s,j} - \mu_{0,j})^2}{n_s}$$

$$\geq \frac{1}{n_0} \left( 1 - \frac{2 \rho_{0,s} \sigma_{s,j}}{\sqrt{n_0 n_s}} + (\mu_{s,j} - \mu_{0,j})^2 \right)$$

$$\geq \frac{(\mu_{s,j} - \mu_{0,j})^2}{n_0} + \frac{1}{n_s} - \frac{1}{n_0};$$

$$w \geq \frac{1}{n_0} + \frac{1}{n_s} - \frac{2 \rho_{0,s} \sigma_{s,j}}{\sqrt{n_0 n_s}} + (\mu_{s,j} - \mu_{0,j})^2.$$


