Regional and temporal patterns of influenza: Application of functional data analysis

Azizur Rahman [a, b], Depeng Jiang [a, c, *]

[a] Department of Community Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada
[b] Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh
[c] School of Sciences, Nanjing Forest University, Nanjing, Jiangsu, China

ABSTRACT

Background: The accurate estimation of temporal patterns of influenza may help in utilizing hospital resources and guiding influenza surveillance. This paper proposes functional data analysis (FDA) to improve the prediction of temporal patterns of influenza.

Methods: We illustrate FDA methods using the weekly Influenza-like Illness (ILI) activity level data from the U.S. We propose to use the Fourier basis function for transforming discrete weekly data to the smoothed functional ILI activities. Functional analysis of variance (FANOVA) is used to examine the regional differences in temporal patterns and the impact of state's political orientation.

Results: The ILI activity has a very distinct peak at the beginning and end of the year. There are significant differences in average level of ILI activities among geographic regions. However, the temporal patterns in terms of the peak and flat time are quite consistent across regions. The geographic and temporal patterns of ILI activities also depend on the political make-up of states. The states affiliated with Republicans had higher ILI activities than those affiliated with Democrats across the whole year. The influence of political party affiliation on temporal pattern is quite different among geographic regions.

Conclusions: Functional data analysis can help us to reveal the temporal variability in average ILI levels, rate of change in ILI levels, and the effect of geographical regions. Consideration should be given to wider application of FDA to generate more accurate estimates in public health and biomedical research.

© 2021 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Influenza is an infectious disease that spreads quickly from one person to another by coughing, sneezing, etc. Influenza is a major cause of morbidity and mortality and remains a serious worldwide threat to public health. Nearly 250,000 to 500,000 deaths occur due to influenza disease worldwide every year (WHO, 2015). As an ongoing project to reduce mortality due to

influenza, the World Health Organization (WHO) has taken a global influenza reduction strategy for the period of 2019—2030 (World Health Organization, 2019).

Influenza usually presents an annual seasonal epidemic or sometimes pandemic. Yearly flu season in North America starts on the 40th week of each year (1st week of October). Since October 1997, the CDC (Center for Disease Control and Prevention, USA) has systematically collected data on the flu season through various channels. Ever since, the CDC has routinely published weekly unrevised unweighted Influenza-like Illness (ILI) activity level data, which measures the percentage of out-patients seeking medical attention for ILI symptoms. The real-time period report is not published by the CDC but rather their most recent ILI report is the ILI activity records of previous weeks. According to the CDC, a person having 100 °F (37.8 °C) or greater body temperature with a cough and/or a sore throat without a known cause other than influenza are considered as symptoms for ILI. The CDC defines ten different levels of ILI activities according to the number of standard deviations below, at or above the mean for the current week compared with the mean of the non-influenza weeks (for details, see: https://www.cdc.gov/flu/index.htm).

To provide recommendations on relevant prevention strategies in the general population, the ILI data compiled by the CDC have been widely analyzed using several different statistical methods to explore the specific features of these data. Some of the methodologies and techniques used so far include, but are not limited to, empirical Bayes (Brooks, Farrow, Hyun, Tibshirani, & Rosenfeild, 2015), SEIR model (Shaman, Karspeck, Yang, Tamerius, & Lipsitch, 2013; Shaman & Karspeck, 2012), linear regression (Yang, Karspeck, & Shaman, 2014), ensemble models with several data sources (Shaman & Kandula, 2015), ARGO model with ensemble of Electronic Health Records and Internet Search Information datasets (Yang et al., 2017), a real-time spatial predictive model (Suhaila & Yusop, 2017; Luca, 2018). All these models treated the time series ILI activities as discrete data. In this paper, we propose to use the recently developed functional data analysis (FDA) models introduced in Ramsay and Silverman (2005) to deal with the continuous curves (called functions or functional data) of ILI data. Generally, ILI activities are observed sequentially over a period (i.e., week, month, or year). Therefore, our data set is time-series in nature (collected weekly for over the year) and our variable of interest (percent unweighted ILI activity levels) can take values over the infinite-dimensional space.

As a non-parametric approach, functional data analysis (FDA) does not make any parametric assumptions about time effects. The basic idea behind FDA is to express discrete time-series observations in the form of a function (to create functional data) that represents the entire measured function as a single observation, and then apply statistical modeling to a collection of functional data. The advantage of doing analysis on functional data is that it can generate models or predictions that can be described by continuous smooth dynamics, which will lead to more accurate estimates of parameters and more effective data noise reduction through curve smoothing. We can also extract additional information from functional data such as the underlying smooth functions, derivatives, and primitives (Ferraty, Mas, & Vieu, 2007; Mas & Pumo, 2009). Another advantage of functional data analysis is that different multivariate procedures such as principal component analysis, discriminant analysis, clustering, and ANOVA can be implemented in the framework of functional data (Peter & Grace, 1978).

The FDA approach of initially smoothing the data and then using the smoothed observations for modeling and forecasting is a major methodological improvement over methods that fit linear/non-linear trends to observed data. These FDA approaches are very suitable for the discovery of temporal patterns in public health (Ullah & Finch, 2013). Somewhat surprising, the use of FDA in public health application has been limited to date. The first objective of the present study is to illustrate how to use FDA to describe basic features of ILI activity levels and to examine the regional difference in temporal patterns of ILI activities.

Under the U.S Constitution, state governors have the authority to take necessary action in public health emergencies. For example, in January 2018, the Alabama state governor declared state of emergency from influenza (flu) and announced travel restrictions. Political orientation (Republican/Democrat and conservative/liberal) and political environment (geo-spatial political party affiliated voting patterns) are both associated with various health outcomes. Several recent studies on the COVID-19 pandemic identified that governors' political affiliation or partisanship acted as an important predictor of the state mandating preventive measures of disease (Adolph, Amano, Bang-Jensen, Fullman, & Wilkerson, 2020; Baccini, Brodeur, & Weymauth, 2021; Clinton, Cohen, & Lapinski, 2020; Neelon, Mutiso, Muller, Pearce, & Neelon, 2020). To date, there has been limited research on examining how the temporal patterns of ILI activities are associated with a state's political orientation. The state's political orientation is defined by the governors' political affiliation (Red or Blue). Another objective of our study is to examine whether and how the geographic and temporal patterns of ILI activities depend on a state's political make-up.

2. Material and methods

2.1. Basics of functional data

Assume that we have a real-valued random process $Y(t)$, which can be expressed as:

$$y_i(t) = x_i(t) + \varepsilon_i(t) \text{ for } i = 1, 2, \cdots, n; t = 1, 2, \cdots, T \tag{1}$$

where $y_i(t)$ was the observed outcome data (i.e., ILI activity level) at the time point t for state i. In our case, the observation space was a closed time interval $[t_{min}, t_{max}]$, and the variable was observed at different time points within the specified time-

space, which provided an observation of the functional variable $y(t_1), \cdots, y(t_T)$. Therefore, we had a dataset that provides n independent and identically distributed functional variable values $y_1, \cdots, y_n$, observed at time points $t_1, \cdots, t_T$ and defined as in equation (1). Here, $T = 52$weeks and $n = 53$states/territories. Each state/territory represented samples (curves or functions), and each week designated time points. $X(t)$ was the underlying continuous process that we wanted to determine and $\varepsilon(t)$ was the independent measurement error of the sample at different time points. In this paper, we considered a Fourier basis function for transforming discrete data into a functional object, and smoothed functional ILI activity levels using penalized smoothing. Generalized cross-validation (GCV) was used to determine the smoothing parameter.

FDA tries to find a smooth function, or better to say, a linear combination of basis functions, that best describes the data recorded at discrete time points. We replaced observations $y_i(t)$ with $x(t_1), \cdots, x(t_T)$, where $x_i(t)$ was the function formed by a linear combination of basis functions $\varnothing_k$ as $x_i(t) \approx \sum_{k=1}^{K} c_k \, \varnothing_k(t)$, here K was the number of bases required to accurately represents the functional data. Here, the conversion of the data into a functional data object required estimation of the expansion coefficients c_k. The choice of suitable basis function depends on the nature of the data set, such as periodic or not periodic. Out of several basis functions, B-spline and Fourier series are the most common.

As our data series collected over time shows periodic nature, the appropriate basis function is Fourier series expansion. The choice of the Fourier basis is also due to high-speed computing and flexibility properties that fits periodic data. The Fourier basis always consists of an odd number of functions. For instance, the first one is the constant basis function, and the remainders are pair of sine and cosine basis functions. Three basis functions are used to generate one harmonic function; two harmonic functions can be obtained from five basis functions, and so on.

Hence, the Fourier series expansion in this paper was written in the form of sine and cosine functions:

$$x_i(t) = c_0 + c_1 \sin(\omega t) + c_2 \cos(\omega t) + c_3 \sin(2\omega t) + c_4 \cos(2\omega t) + \cdots \tag{2}$$

Equation (2) was defined by the basis $\varnothing_1(t) = 1, \varnothing_{2k-1}(t) = \sin(k\omega t), \varnothing_{2k}(t) = \cos(k\omega t)$ with $t = t_1, \cdots, t_T$. The periodic basis was considered and ω was found in relation with the period T by the formula: $\omega = 2\pi/T$. The expansion coefficients c_k can be obtained by minimizing the least-squares criterion expressed in matrix notation as

$$SSE = (\boldsymbol{y} - \varnothing\boldsymbol{c})'(\boldsymbol{y} - \varnothing\boldsymbol{c}) \tag{3}$$

Here, the matrix $\varnothing$ was the value of basis functions $\varnothing_k(t)$, the vector $\boldsymbol{c}$ referred to expansion coefficients and $\boldsymbol{y}$ represented the vector of observations at time t. The estimated value of the basis coefficient was

$$\widehat{\boldsymbol{c}} = (\hat{\varnothing}'\hat{\varnothing})^{-1}\hat{\varnothing}'\boldsymbol{y} \tag{4}$$

The general approach of representing periodic data is to consider a small number of basis functions. It is recommended that the roughness penalty approach could be used while using many basis functions. The roughness penalty is intended to smooth the function and reduce overfitting. The penalized sum of square of error (PENSSE) is used as a criterion to estimate the expansion coefficient vector $\boldsymbol{c}$ in order to minimize errors of curves (Peter & Grace, 1978; Chebana, Daobo-Niang, & Ouarda, 2012) as in equation (5).

$$PENSSE_\lambda(\boldsymbol{x}) = \sum_j \left[\boldsymbol{y}(\boldsymbol{t_j}) - \boldsymbol{x}(\boldsymbol{t_j})\right]^2 + \lambda \int \left[\boldsymbol{D}^2\boldsymbol{x}(\boldsymbol{t})\right]^2 dt \tag{5}$$

where, $\boldsymbol{x}(\boldsymbol{t})$ defined in equation (2); the second term of above equation includes the roughness penalty, which calculates the overall curvature of the second derivatives of function $\boldsymbol{x}(\boldsymbol{t})$. The smoothing parameter λ measures the roughness of the fitted curve, and controls the trades off curve roughness against lack of data fit (Ramsay, Hooker, & Graves, 2009). The small value of λ determines the better fit of $\boldsymbol{x}(\boldsymbol{t})$ to the data, whereas large value makes $\boldsymbol{x}(\boldsymbol{t})$ into linear. Therefore, the vector of fits to the data according to Ramsay and Silverman (Ramsay & Silverman, 2005) is:

$$\boldsymbol{S}_\varnothing\boldsymbol{y} = \varnothing(\varnothing'\boldsymbol{W}\varnothing + \lambda\boldsymbol{R})^{-1} \varnothing'\boldsymbol{W}\boldsymbol{y} \tag{6}$$

where, $\boldsymbol{S}_\varnothing$ is a projection operator for each of basis functions $\varnothing$, $\boldsymbol{y}$ is vector of data to be smoothed, $\boldsymbol{R}$ is known as penalty matrix which can be obtained by second derivative of $\varnothing$ defined as $\boldsymbol{R} = \Delta'\Delta$ in Yang et al. (2017):

$$\Delta = \begin{bmatrix} 0 & \omega^2\sin(\omega t_1)\cdots & \cdots k^2\omega^2\cos(\omega t_1) \\ \vdots & \vdots & \vdots \\ 0 & \omega^2\sin(\omega t_T)\cdots & \cdots k^2\omega^2\cos(\omega t_T) \end{bmatrix}$$

The smoothing parameter λ was chosen by generalized cross validation (GCV) criterion as introduced by Mas and Pumo (2009) and is given by:

$$GCV_\lambda = \frac{n^{-1} \sum [\boldsymbol{y}(t_j) - \boldsymbol{x}(\boldsymbol{t_j})]^2}{[n^{-1} \ trace \ (\boldsymbol{I} - \boldsymbol{S})]^2} \tag{7}$$

and we defined the number of degrees for smooth fit as $df(\lambda) = trace \ (\boldsymbol{S}_{\varnothing,\lambda})$. Here, the best choice of λ was determined with the minimum value of GCV.

2.2. Descriptive analysis with functional data

We can define the mean and variance of functional data, similarly to traditional descriptive statistics. The mean and variance of functional data are given in (8) and (9), respectively.

$$\mu(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t) \tag{8}$$

$$VAR_X(t) = \frac{1}{N-1} \sum_{i=1}^{N-1} [x_i(t) - \overline{x}(t)]^2 \tag{9}$$

We know by definition that the standard derivation was represented by positive square root of equation (9). The dependence measure of curves in functional data analysis is represented by covariance functions between curve values. Suppose that we represent the two curves $x_i(s)$ and $x_i(t)$ at times s and t. Then the dependence between two curves can be expressed by sample functional covariance as defined in Ramsay et al. (2009)

$$Cov(s,t) = v(s,t) = \frac{1}{N-1} \sum_{i=1}^{N-1} (x_i(s) - \overline{x}(s))(x_i(t) - \overline{x}(t)) \tag{10}$$

The variability of the functional sample can be visualized by plotting the surface of covariance as functions of s and t along with a corresponding contour map. In addition, functional principal component analysis is recommended for describing variation in functional curve from their mean (for detailed discussion, See Ramsay et al., 2009). In multivariate settings, it is known that the eigen analysis of the covariance matrix $\boldsymbol{V}$, solving the matrix eigen equation $\boldsymbol{V} \Psi_j = \mu_j \Psi_j$, would produce the whole suite of eigenvalue/eigenvector pairings. For functional data, the procedure is roughly the same; we calculate eigenfunctions Ψ_j of the bivariate covariance function $\mathbf{v}(s,t)$ as solutions of the functional eigenequation using equation (11)

$$\int v(s,t)\Psi_j(t)dt = \mu_j \Psi_j(s) \tag{11}$$

2.3. Analysis of variance of functional data

A natural statistical problem in FDA is to decide about the existence of differences between curves obtained under different levels of some categorical factors. The analogous idea of classical ANOVA was used for functional objects in which mean functions for each group are considered for comparison purposes. This approach is called functional ANOVA (FANOVA). The advantage of using Functional ANOVA over classical ANOVA is that it can use all the information in the curve rather some specific values. Computational procedures of FANOVA are almost identical to classical ANOVA except for some changes on it.

In this paper, we used the FANOVA technique introduced by Ramsay and Silverman under the framework of functional regression analysis (Ramsay & Silverman, 2005). The response was a set of functions or curves, and the covariate was a set of dummy variables based on the categorical factors. Suppose that there are G different groups of a categorical factor which may affect the random process $Y(t)$ and $Z^1, \cdots, Z^G$ represents the dummy indicators for each group. Therefore, the one-way FANOVA is represented by the following functional regression:

$$y_{\boldsymbol{i}}(t) = \beta_0(t) + \sum_{g=1}^{G-1} \beta_g Z_g + \varepsilon_i(t) \tag{12}$$

here, regression coefficients were considered as time-varying functions, and hence it is also known as the varying coefficient model. At each time t, the defined functional linear regression model in matrix notation is:

$$\boldsymbol{y_i}(t) = \boldsymbol{U_i} \beta(t) + \varepsilon_i(t) \tag{13}$$

where $\boldsymbol{U}$ represents the design matrix. Similarly, as in Ordinary Least Square (OLS) regression, the estimated regression coefficients can be obtained by minimizing the residual sum of squares (SSE). The estimates for regression coefficients can be defined in matrix notation as below

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{t}) = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Y}(\boldsymbol{t}) \tag{14}$$

Our approach was to obtain the smooth function of $\widehat{\boldsymbol{\beta}}(\boldsymbol{t})$ by the linear combination of the basis functions $\varnothing_k$. The smoothed coefficient is defined as $\widehat{\boldsymbol{\beta}}(\boldsymbol{t}) \approx \sum\limits_{m=1}^{p} c_m \varnothing_m(t)$. Therefore, our smooth least square estimate is

$$\widehat{\boldsymbol{b}} = \left[\sum \int \psi_i(t) \psi_i(t)' dt\right]^{-1} \left[\sum \int \psi_i(t)' y_i(t) dt\right] \tag{15}$$

where, $b = [c_1', \cdots, c_p']'$ and $\psi_i(t) = [z_{i1}\varnothing_1(t), \cdots, z_{ip}\varnothing_p(t)]$. Hence, the test of significance for our overall difference between the groups was obtained by permutation test procedure since, in functional setup, the distribution of the test statistic under null hypothesis was quite complicated and difficult to obtain. The null hypothesis for group effect was $H_0 : \beta_l = 0 ; l = 1, \cdots, G-1$. The test statistic was based on $F^* = \max F(t)$, where $F(t)$ is pointwise F statistic, defined as

$$F(t) = \frac{Var\left(\widehat{\boldsymbol{y}}(\boldsymbol{t})\right)}{\sum\left(y_i(t) - \widehat{y}_i(t)\right)^2} \tag{16}$$

Similarly, two-way FANOVA can be conducted in functional regression framework as above.

3. Results

This study analyzed the U.S weekly influenza data from the year 2011−2019 reported by the CDC. CDC measures influenza-like illness (ILI) through percentage of out-patients visits with ILI symptoms. The dataset consists of weekly measurements of ILI activity levels for each state/territory. The 53 states/territories were divided into five regions: Northeast (consists of 9 states), Midwest (consists of 12 states), South (consists of 17 states), West (consists of 13 states) and territories group (includes Puertro Rico, Virgin Islands and New York City).

First, we transformed our weekly ILI values into a functional object. As our data shows periodicity, we represented our discrete data with Fourier basis functions. According to the generalized cross validation (GCV) criteria, we found that 15 Fourier basis functions would be an appropriate choice to represent our data as functional data objects.

Next, we obtained the best estimated curves $\boldsymbol{x}(\boldsymbol{t})$ by eliminating the contribution of the errors and noise presented in the functional objects. We chose generalized cross-validation (GCV) criterion to get the level of smoothness by placing the harmonic acceleration operator. Here the linear differential operator (LDO) derived by the formula $(2\pi/52)^2$ was used, and the smoothing parameter was chosen based on minimization criterion. The best choice of the smoothing parameter estimated holds the smoothing properties, which could reflect the data well and capture a high percentage of explained variance. Hence,
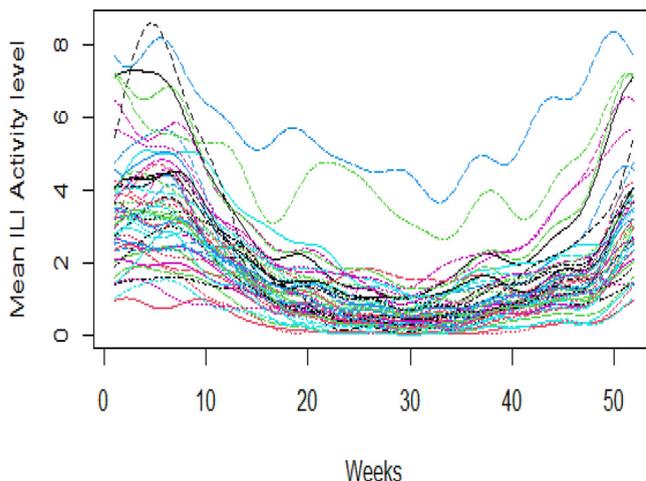

Fig. 1. Plot of the smooth functional observations for mean weekly ILI activities.

the corresponding smoothing parameter and degrees of freedom for accurately represent our dataset was 0.04 and 15, respectively. The smoothed average ILI activity levels curves for all years for all 53 states/territories are shown in Fig. 1. Further analyses of geographical and temporal patterns are based on these functional data.

3.1. Temporal patterns of influenza

Fig. 2 shows the mean and standard deviation of the functional data across states and territories. The highest peak was between the 6th to 8th week's period and the next highest value observed on the 48th to 50th period of weeks. The high variation was also found at both periods compared to other points of time. On the other hand, the lowest value observed in the mean function at the 30th week point of time, and so low variation was also found at this time point.

Fig. 3 shows the rate of change in the mean function of average ILI activity levels, which is not available from the traditional statistical analysis. A positive change rate in mean ILI activity levels was observed in the first few weeks (6th to 8th) and last few weeks (48th to 50th) of the year, while the negative rates were shown at the first few weeks of the year and appeared relatively stable in the middle of the years. Fig. 4 shows two rotated principal component functions or harmonics as perturbations of the mean. The left panel of Fig. 4 contains the strongest component (explains 53.2% variations), indicating primarily the highest variation around week 22. The right panel of Fig. 4 (explains 42.2% variations) contains the second principal component functions, indicating the highest variation in the first few weeks.

3.2. Regional differences in temporal patterns of influenza

One-way FANOVA was conducted to examine the regional differences in temporal patterns of ILI activities. The results are shown in Fig. 5, with the top left panel showing the national average ILI activity levels and the remaining panels showing how each region differs from the national average. The mean ILI activity in the South region was above the national average over all weeks, whereas the mean ILI activity in territories was below the national average. For most weeks in a year, the mean ILI activity for the Midwest region was below the national level, whereas the mean ILI activity for the West was above the national level. For the Northeast region, the mean ILI activity was above the national average in first and last few weeks and middle of the year, whereas for the other weeks was lower than the national average.

The result of the permutation F-test in FANOVA is shown in Fig. 6. The observed statistic curve was below the maximum 0.05 critical values at each time, indicating that the overall regional differences in temporal pattern were not statistically significant. The observed F statistic curve was above the pointwise 0.05 critical value during Week 46 to Week 50, indicating pointwise regional differences exist between this time-period, which is statistically significant.

3.3. Political orientation and temporal patterns of influenza

To examine how the temporal patterns of ILI activities are associated with a state's political orientation, we conducted a one-way FANOVA on ILI activities with state's political orientation as a categorical factor. Fig. 7 shows how red states (top panel) and blue states (bottom panel) differ from the national average. The red states had higher ILI activities than the national average whereas the blues states had lower levels across the whole year. The permutation F-test, shown in Fig. 8, shows that the observed statistic curve was above the maximum 0.05 critical values all the year, indicating that the impact of political orientation on the temporal pattern was statistically significant at each time-period. Fig. 7 indicates that the highest positive effect of red states was attained from first week to 5th week period and between the 48th week to 50th week period, whereas the highest negative effects of blue states were revealed from week 3 to week 8 and from week 45 to week 50. These results indicate that the impacts of political orientation on ILI activities are more dominant during the flu transmission period.
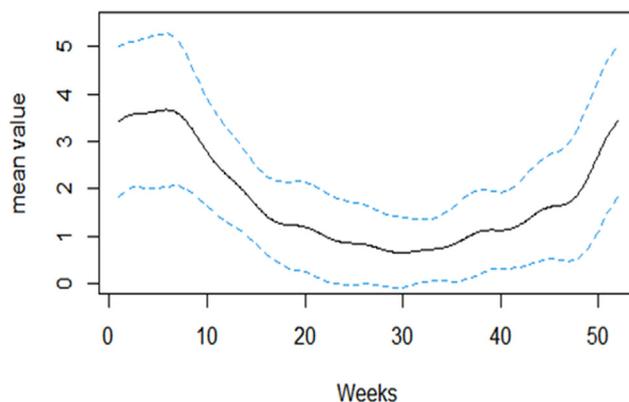

Fig. 2. Mean and standard deviation of functional data. Solid line represents mean and dotted lines represent one standard deviation above or below mean.
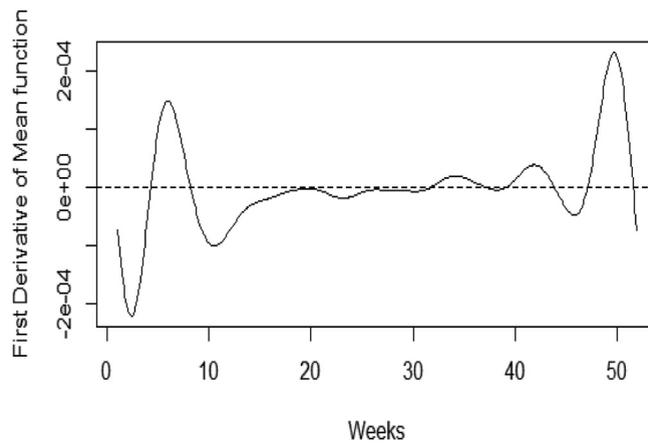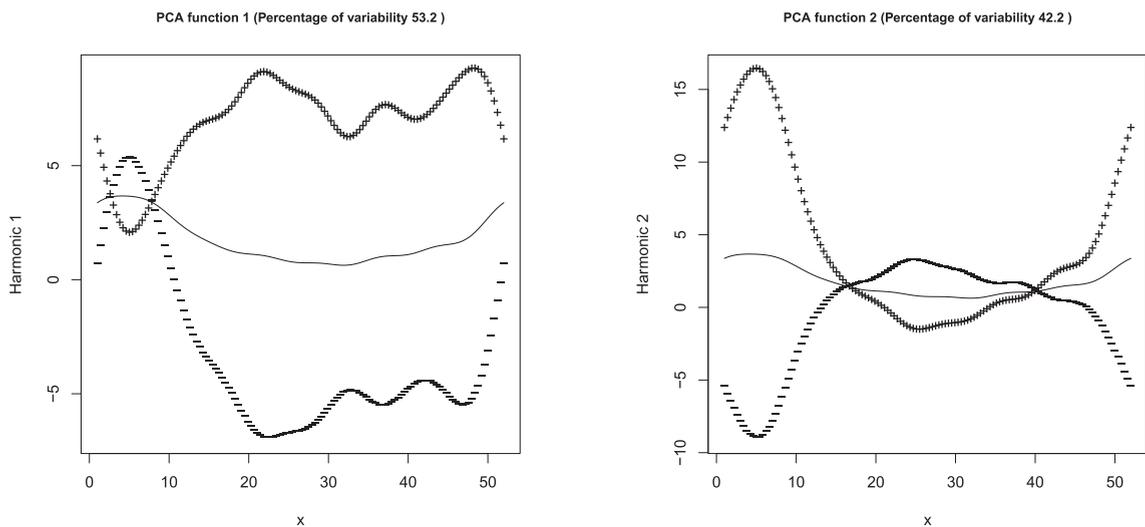
Fig. 3. Rate of change in mean functions.

Fig. 4. Two principal component functions as perturbations of mean. Left panel shows the first principal component and right panel shows the second principal component.
Note: The two rotated principal component functions or harmonics are shown as perturbations of the mean, which is the solid line. The +'s show what happens when a small amount of a principal component is added to the mean, and the -'s show the effect of subtracting the component.

To examine whether the regional difference in temporal patterns of ILI activities depends on political orientation, we conducted two-way FANOVA with state's political orientation and the region as categorical factors. Fig. 9 shows the mean weekly ILI activities in each region and political states. In both Midwest and Northeast regions, the temporal patterns are similar between red and blue states for most time periods. In the West region, the average ILI activities are higher in blue states than in red states, whereas in the South region, the average ILI activities are more elevated in red states than in blue states for the most period. Fig. 10 provides point-wise F statistic results with permutation test to test the significance of the interaction effects using functional regression model. It is clearly depicted that observed or calculated F statistic curve cross the maximum pointwise 0.05 critical value line at the end of year starting from week 46 to week 49, indicating the presence of significant interaction effect between regions and political orientation.

4. Discussions and conclusions

This paper introduces a novel idea for temporal patterns of infectious disease: using functional data analysis technique and comparing the predicted continuous distribution rather than observed discrete values. Using the U.S. ILI activity data, we have considered the Fourier basis function for transforming discrete weekly data to the smoothed functional ILI activity levels. The plot of smoothing curves for average weekly ILI activity levels can be used to identify the peak and flat seasons for each region.
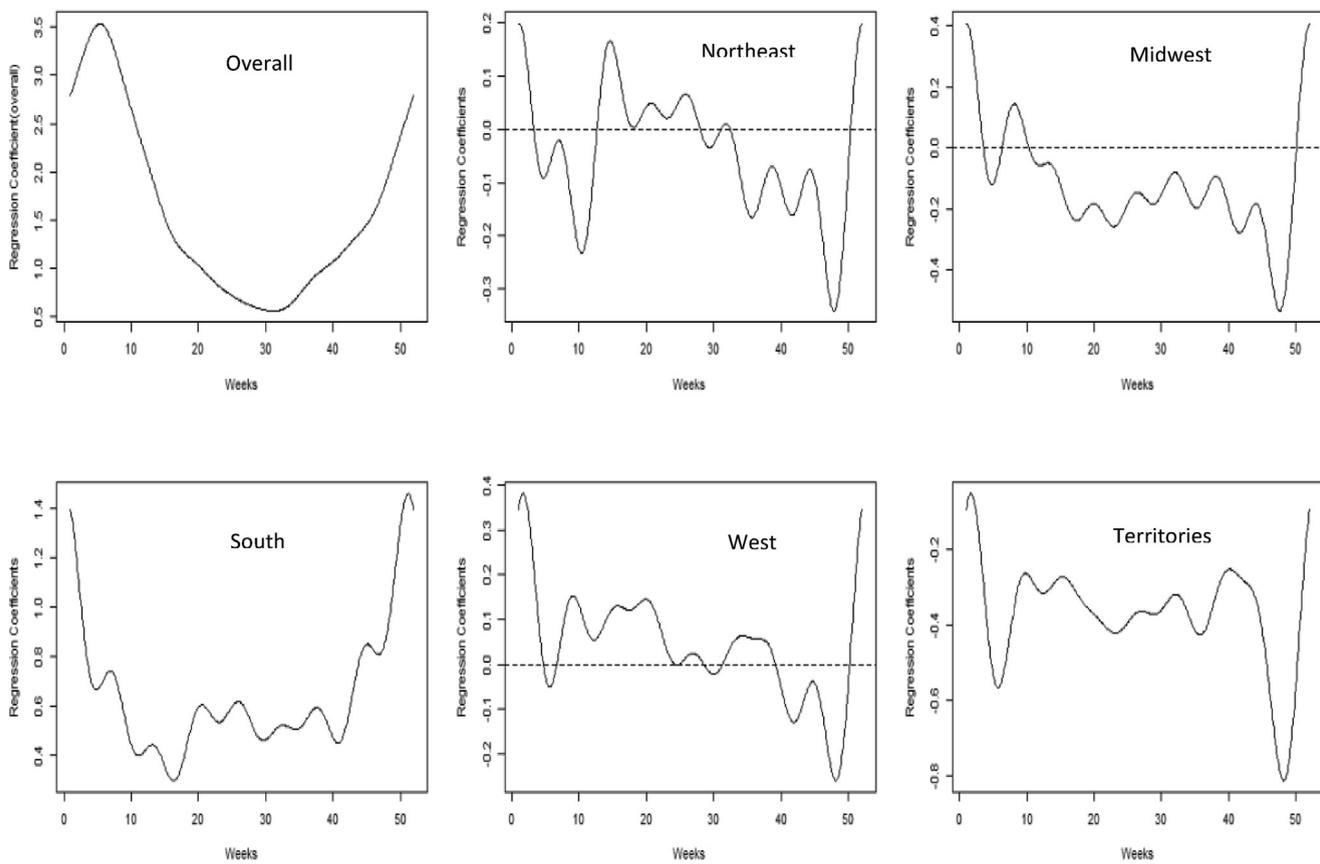
Fig. 5. The overall temporal pattern of ILI activities and regional deviations from the overall patterns.
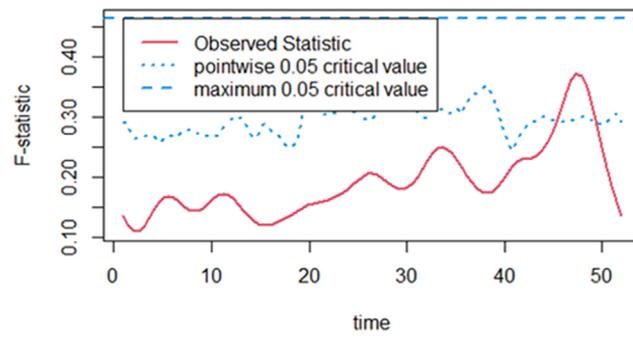
Fig. 6. The permutation F-test for the regional differences in functional ILI levels.
Note: Solid line (red color): observed F-statistic value; Dotted line maximum pointwise: 5% critical value; Dashed line: permutation 5% critical value.
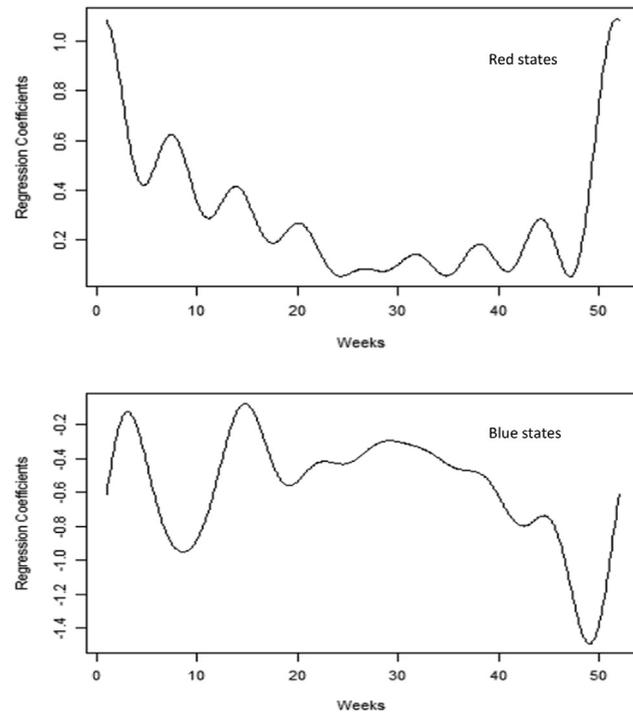

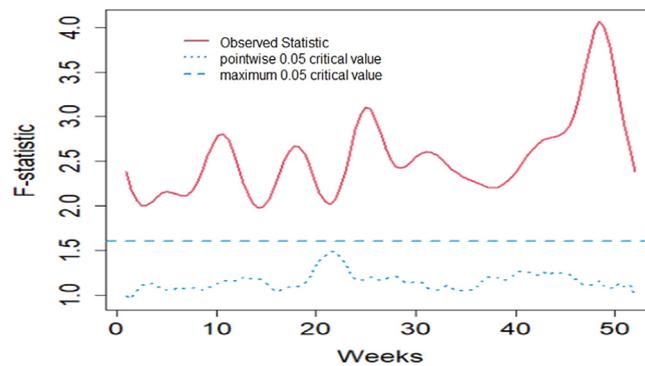**Fig. 7.** Effects of political orientation on temporal patterns of influenza.

Fig. 8. Permuted F-test of differences in ILI activity levels for political orientations.
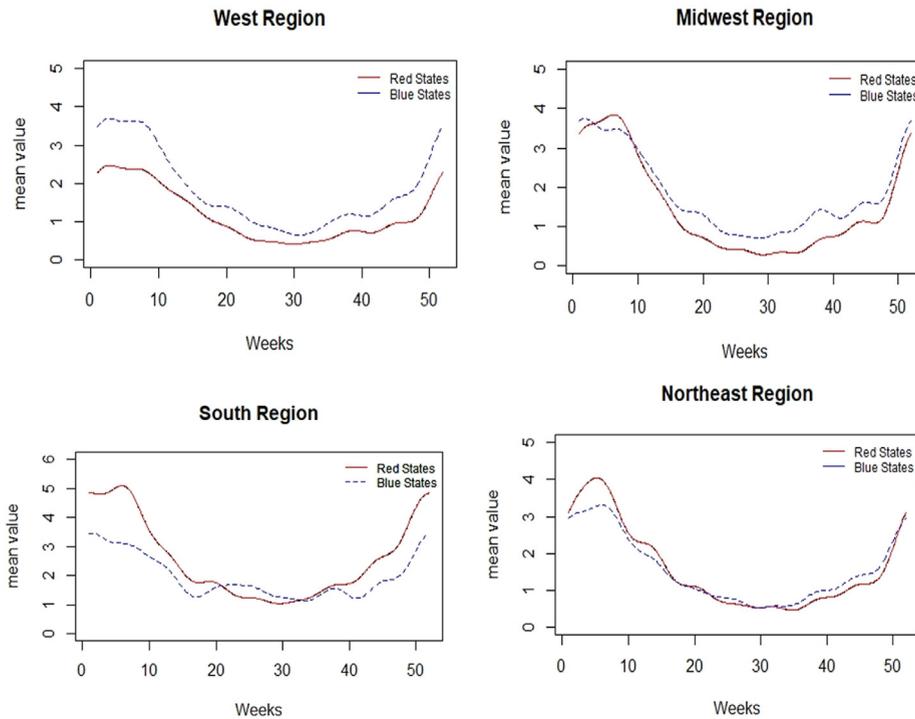
Fig. 9. Impact of political orientation on temporal patterns of influenza by region.
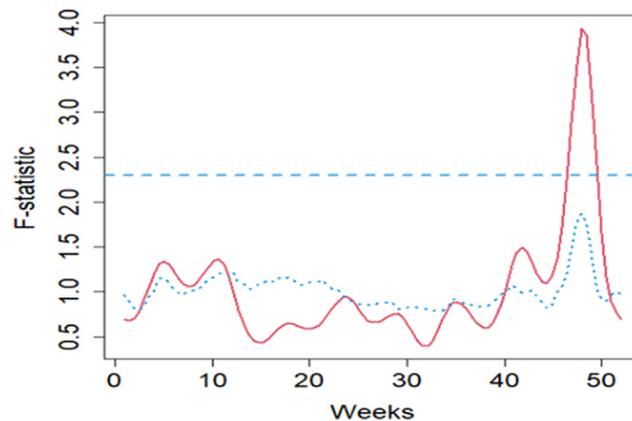

Fig. 10. The permutation F-test for interaction effect of political states and regions.
Note: The solid line (red color) represents observed F-statistic value, the dotted line maximum pointwise 5% critical value and the dashed line represents permutation 5% critical value.

We have also shown how to use functional ANOVA or functional regression to examine factors associated with temporal patterns of ILI activities.

Our results reveal that the ILI activity has a distinct peak at the beginning and end of the year, more specifically during week 1—6 and week 48—51, while it is flatter over other periods. We also found the high increase rate of ILI activities in the first few weeks (6th to 8th) and the last few weeks (48th to 50th) of the year, while negative rates were shown at the first few weeks of the year and were relatively stable in the middle of the years. This temporal pattern of change rate of ILI activities could not be discovered by analyzing the observed discrete data.

The result of functional ANOVA indicates significant differences in average level of ILI activities among geographic regions. The South region has a higher average ILI level than the national average, and the territories has lower level across all time-periods. However, the temporal pattern in terms of the peak and flat time are quite consistent across regions as the point-wise permutation test in functional ANOVA does not reveal statistical significance.

Our results confirm that the political environment is associated with temporal pattern of ILI activities. The state and county political environment impacts resident's flu vaccination behaviors in that Republican individuals are less likely to receive flu vaccination. The states affiliated with Republicans had higher ILI activities than those affiliated with Democrats across the whole year. An exciting aspect of our results is that the influence of political party affiliation on the temporal pattern is quite different among geographic regions. For example, in the West region, the average ILI activities are higher in blue states than in red states, whereas in the South region, the average ILI activities are higher in red states than in blue states. This might be related that the residents of Republican states (presumably, living under conservative policies) are more likely to get the flu vaccine. Understanding health behavior adoption at both the individual, ecological and political levels could inform strategies in crafting and advocating for the public health policy.

For future research, it would be interesting to take into account some additional available information (e.g., temperature and climate changes) in the functional regression on prediction of ILI activity levels. The functional analysis technique could be adapted for longitudinal setting to consider an unusual year (like a pandemic year) and to assess the impact of a pandemic outbreak on temporal patterns of influenza.

It is imperative to understand temporal patterns of infectious disease in public health research. It is critical that such pattern discovery is based on the best available statistical modeling approaches to minimize possible prediction errors. Functional data analysis can help us to reveal the temporal variability in average ILI levels, rate of change in ILI levels, and geographical regions' effect. Our results can provide additional insight for influenza surveillance strategies, hospital resource management, and influenza outbreaks measurement. Consideration should be given to the wider application of functional data analysis (FDA) to generate more accurate estimates in public health and biomedical research.

Declaration of competing interest

All authors have no conflicts of interest to disclose.

Acknowledgement

References

Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., & Wilkerson, J. (2020). Pandemic politics: Timing state-level social distancing responses to COVID-19. *Journal of health politics, policy and law.* https://doi.org/10.1215/03616878-8802162

Baccini, L., Brodeur, A., & Weymauth, S. (2021). The COVID-19 pandemic and the 2020 U.S. Presidential election. *Journal of Population Economics, 34*, 739–767. https://doi.org/10.1007/s00148-020-00820-3

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., & Rosenfeild, R. (2015). Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Computational Biology, 11*, Article e1004382. https://doi.org/10.1371/journal.pcbi.1004382

Chebana, F. S., Daobo-Niang, S., & Ouarda, T. B. (2012). Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research, 48*(4). https://doi.org/10.1029/2011WR011040

Clinton, J., Cohen, J., Lapinski, J., & Trusster, M. (2020). Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19. *Science Advances.* , Article eabd7204. https://doi.org/10.1126/sciadv.abd7204

Ferraty, F., Mas, A., & Vieu, P. (2007). Advances in nonparametric regression for functional variables. *Australian & New Zealand Journal of Statistics, 49*(3), 267–286. https://hal.archives-ouvertes.fr/hal-00635687v2.

Luca, B., Manuel, O. F., Nuria, T., Ana, M., & Mireia, J. (2018). Real-time predictive seasonal influenza model in Catalonia, Spain. *PloS One, 13*(3), Article e0193651. https://doi.org/10.1371/journal.pone.0193651

Mas, A., & Pumo, B. (2009). Functional linear regression with derivatives. *Journal of Nonparametric Statistics, 21*, 19–40. https://doi.org/10.1080/10485250802401046

Neelon, B., Mutiso, F., Muller, N. T., Pearce, J. L., & Neelon, S. B. E. (2020). Associations between governor political affiliation and COVID-19 cases and deaths in the United States. *medRxiv.* https://doi.org/10.1101/2020.10.08.20209619

Peter, C., & Grace, W. (1978). Smoothing noisy data with spline functions: Estimating the current degrees of smoothing by the method of generalized cross-validation. *Numerische Mathematik, 31*, 377–403. https://doi.org/10.1007/bf01404567

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB* (1ˢᵗ ed.). New York: Springer Series in Statistics. ISBN: 978-0-387-98184-0.

Ramsay, J. O., & Silverman, B. (2005). *Functional data analysis*. New York: Springer Series in Statistics. ISBN: 978-0-387-22751-1.

Shaman, J., & Kandula, S. (2015). Improved discrimination of influenza forecast accuracy using consecutive predictions. *PLoS Current Outbreak, 1*(4). https://doi.org/10.1371/currents.outbreaks.8a6a3df285af7ca973fab4b22e10911e

Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. In *Proceeding of national academy of science, U S A.* (Vol. 109, pp. 20425–20430).

Shaman, J., Karspeck, A., Yang, W., Tamerius, J., & Lipsitch, M. (2013). Real-time influenza forecasts during the 2012-2013 season. *Nature Communications, 4*, 2837.

Suhaila, J., & Yusop, Z. (2017). Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology, 129*, 229–242. https://doi.org/10.1007/s00704-016-1778-x

Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology, 13*, 43. http://www.biomedcentral.com/1471-2288/13/43.

World Health Organization. (2015). Fact sheet number 211. Available from http://www.who.int/mediacentre/factsheets/fs211/en/index.html. (Accessed 7 July 2020).

World Health Organization. (2019). Global influenza strategy. available from https://www.who.int/influenza/global_influenza_strategy_2019_2030/en. (Accessed 7 July 2020).

Yang, W., Karspeck, A., & Shaman, J. (2014). Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Computational Biology, 10*, Article e1003583. https://doi.org/10.1371/journal.pcbi.1003583

Yang, S., Santillan, M., Brownstein, J. S., Gray, J., Richardson, S., & Kou, S. C. (2017). Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infectious Diseases, 17*, 332. https://doi.org/10.1186/s12879-017-2424-7