Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(8): 6936–6974.
DOI: 10.3934/mbe.2024305
Received: 04 April 2024
Revised: 16 July 2024
Accepted: 24 July 2024
Published: 27 August 2024

Research article

Analyzing factors of daily travel distances in Japan during the COVID-19 pandemic

Masaya Mori, Yuto Omae, Yohei Kakimoto, Makoto Sasaki and Jun Toyotani*

College of Industrial Technology, Nihon University, Izumi, Narashino, Chiba, Japan

* **Correspondence:** Email: toyotani.jun@nihon-u.ac.jp.

Abstract: The global impact of the COVID-19 pandemic is widely recognized as a significant concern, with human flow playing a crucial role in its propagation. Consequently, recent research has focused on identifying and analyzing factors that can effectively regulate human flow. However, among the multiple factors that are expected to have an effect, few studies have investigated those that are particularly associated with human flow during the COVID-19 pandemic. In addition, few studies have investigated how regional characteristics and the number of vaccinations for these factors affect human flow. Furthermore, increasing the number of verified cases in countries and regions with insufficient reports is important to generalize conclusions. Therefore, in this study, a group-level analysis was conducted for Narashino City, Chiba Prefecture, Japan, using a human flow prediction model based on machine learning. High-importance groups were subdivided by regional characteristics and the number of vaccinations, and visual and correlation analyses were conducted at the factor level. The findings indicated that tree-based models, especially LightGBM, performed better in terms of prediction. In addition, the cumulative number of vaccinated individuals and the number of newly infected individuals are likely explanatory factors for changes in human flow. The analyses suggested a tendency to move with respect to the number of newly infected individuals in Japan or Tokyo, rather than the number of new infections in the area where they lived when vaccination had not started. With the implementation of vaccination, attention to the number of newly infected individuals in their residential areas may increase. However, after the spread of vaccination, the perception of infection risk may decrease. These findings can contribute to the proposal of new measures for efficiently controlling human flows and determining when to mitigate or reinforce specific measures.

Keywords: machine learning; correlation analysis; COVID-19; human flow; human mobility; Peltzman effect

1. Introduction

Since the confirmation of COVID-19 in 2019, health hazards have increased [1, 2]. Social [3, 4] and economic impacts [5, 6] associated with the spread of the infection have been recognized as global problems. Therefore, various factors related to the spread of COVID-19 infection have been investigated to date, and corresponding preventive measures have been discussed extensively. For example, Coccia investigated the specific conditions contributing to the spread of COVID-19 and found that wind speed and air pollution affect the spread of COVID-19 [7–9]. Bontempi et al. found that international trade activities are a strong factor affecting the spread of COVID-19, possibly due to increased contact among foreigners [10, 11]. Other contributing factors to the spread of infection include urban population density, absolute humidity [12], vaccination [13], and seasonality [14, 15]. These investigations are expected to be applied not only to the suppression of COVID-19, but also to the preparation against the next pandemic similar to COVID-19. Therefore, it is important to investigate and discuss these issues, although the COVID-19 pandemic is being contained.

To control the spread of infection, it is particularly important to identify the factors that influence human flow (that is, human mobility, travel distances, and population density at any given place and time) and to understand their detailed trends. There are four reasons for this:

- Human mobility has been reported to affect COVID-19 transmission more strongly than other factors (population density, temperature, vaccination coverage, etc.) [16], and government policies can intervene.
- In countries and regions with well-developed public transportation systems (trains, buses, airplanes, etc.), people can easily travel for work, entertainment, and other purposes. Thus, human mobility is likely to have a significant impact on the spread of COVID-19 in such countries and regions [17].
- Human flow has been reported to have a strong relationship with the effective reproduction number regardless of the first or second wave of COVID-19 [18], and therefore, a continued effect on the control of infection can be expected.
- Human flow is also highly relevant in other infectious diseases (influenza virus and mosquito-borne infectious diseases such as malaria and dengue fever), and the findings obtained may be widely applicable [19].

In general, human flow is a crucial factor in the transmission of infectious diseases, which spread rapidly with an increase in human flow. Consequently, pandemics are more likely to occur when preventing and controlling infectious diseases is challenging [20]. Therefore, during the early stages of the COVID-19 outbreak, countries implemented travel restrictions, lockdowns, and emergency declarations to limit the flow of people into and out of the country. Such non-pharmaceutical interventions have been reported to be effective in controlling COVID-19 transmission in many countries and regions [21–23]. Following the development of the COVID-19 vaccine and the availability of pharmaceutical interventions, COVID-19 vaccinations became widely available and movement restrictions were eased. COVID-19 vaccination has been effective in reducing the risk of infection [24, 25] and preventing serious illness and death from infection [26–28]. However, a positive correlation between COVID-19 vaccination coverage and increased human flow has been observed worldwide [29, 30]. This has also been reported to reduce the effectiveness of COVID-19 vaccines [31, 32]. Moreover,

approximately 30% of the population is hesitant to vaccinate, and enforcing coercive measures can undermine democratic values and impact the socioeconomic system [33]. Therefore, it is necessary to understand the factors that affect human flow not only when pharmaceutical interventions are not possible but also when they are possible as managing and controlling these factors is important.

It is known that human flow is determined by a complex interplay of multiple factors. Examples include economic factors such as income levels and prices [34], social factors such as population composition, culture, and customs [35], environmental factors such as temperature and weather [36], and commercial event factors such as the Olympics [37,38]. Among these, human flow in the COVID-19 disaster is expected to be strongly influenced by political and legal factors relating to domestic and international movement restrictions as well as by health and safety factors that indicate the prevalence of COVID-19. Therefore, it is important to measure the effects of political and legal factors on human flow and to elucidate the conditions for maximizing these effects. In addition, information on health and safety factors, including statistical data, public data, public health, and infection trends, has been disclosed to the public by the government and mass media. Clarifying the impact of such information on human flow will be helpful for future measures to prevent the spread of infection and to resume economic activities. That is, elucidating the effects of changes in political and legal, as well as health and safety, factors on human flow is essential to consider future countermeasures.

In response to these issues, machine learning has been applied extensively as a human flow analysis method for COVID-19 in recent years. For instance, Hu et al. employed generalized additive mixed models with big data in the US to evaluate the effects of government policies on human mobility, and discovered that such policies had limited, time-decreasing, and region-specific effects on mobility [39]. Nakamoto et al. observed that the state of emergency in Japan deterred human flow, although its effect diminished with subsequent declarations [40]. Chakraborty et al. used regularized linear models to forecast human mobility in the US, analyzing various factors and noting a decrease in the number of daily trips as nine factors, including the number of newly infected individuals, median income, and socioeconomic status, increased [41]. Additional studies indicated that travel restrictions [42] and lockdowns [43] that were imposed early in an outbreak within a specific region can effectively regulate the human flow.

The factors that increase or decrease human flow in a pandemic are gradually being elucidated through machine learning. These studies are expected to be applied to control human flow during the spread of infectious diseases, including COVID-19. However, although numerous studies have aimed to identify information and measures that are effective in regulating human flow, research focusing on information and measures, particularly those associated with human flow, remains limited. Furthermore, few studies have undertaken factor analysis to account for the impact of regional differences relating to published information, such as whether it concerns residential areas close to one's home or areas of socioeconomic activity, including workplaces and schools. In addition, the impact of the frequency of the implemented measures as well as the combined effect of these factors has seldom been considered. Moreover, as these discussions often pertain to specific countries and regions, broader surveys encompassing various locations are essential for generalizing the conclusions. However, conducting verification at all sites is difficult in reality; therefore, it is important to increase the number of reference cases in countries and regions with insufficient verified cases. Thus, it is necessary to conduct research that satisfies the following three requirements: 1) Identify factors that have a particularly high association with human flow in COVID-19. 2) Clarify how regional characteristics and differences in

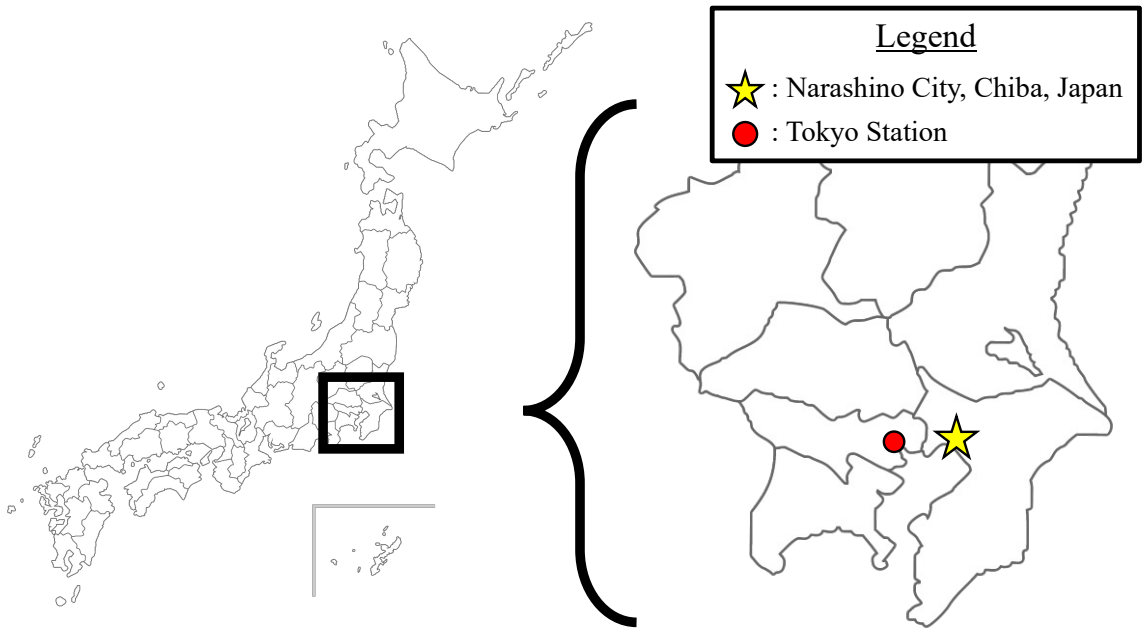


Figure 1. The population area for which the average daily distance traveled was measured.

vaccination of these factors affect changes in human flow. 3) Conduct surveys targeting countries and regions with insufficient verified cases.

Therefore, in this study, the average daily travel distance was defined as the human flow, with the following three objectives: 1) Mainly political and legal factors and health and safety factors that are expected to have a strong impact on human flow in the event of a COVID-19 disaster are analyzed using machine learning. 2) The factors identified from the analysis results are subdivided according to regional characteristics and vaccination, and the detailed effects are confirmed through visual and correlation analyses. 3) Narashino City, Chiba Prefecture, Japan, is targeted as a country or region with insufficient survey cases. The remainder of this paper is organized as follows: Section 2 provides an overview of human flow data as the response variable; an overview of political, legal, health, and safety factors as explanatory variables (features); and a description of the procedures for human flow analysis using machine learning. Section 3 describes the performance evaluation of machine-learning models and comparison results between models, and presents the results of factor analysis using machine learning, visual analysis, and correlation analysis. Section 4 describes the results and discusses the pros and cons of this study. Section 5 summarizes the study and discusses future prospects.

2. Materials and methods

2.1. Sample and data

This section provides an overview of Narashino City, Chiba Prefecture, Japan, which was the target of the study, and describes the human flow data of Narashino residents as the response variable. Figure 1 shows a map of Narashino City, located in the northwestern part of Chiba Prefecture, Japan. As of August 31, 2023, Narashino had a population of 175,260 within a geographical area of 20.97 km^2. The city is within 30 km of Tokyo, with travel times to the Tokyo train station of approximately 45 min

Table 1. Mean ± SD of the number of people and average daily travel distance measured by age group.

Age group	People (number/day)	Average travel distance (m/day)
Teens	92 ± 47	6157 ± 748
Twenties	159 ± 77	6410 ± 568
Thirties	143 ± 67	5845 ± 596
Forties	108 ± 55	6847 ± 680
Fifties	81 ± 42	6225 ± 683
Sixties	24 ± 13	5557 ± 1167
Seventies	6 ± 3	4601 ± 2261
Unknown	394 ± 85	5265 ± 517
Total	1007 ± 355	5867 ± 1240

by train and 35 min by car. In this study, the average daily distance traveled by residents of Narashino, Chiba, was used as the human flow data in the COVID-19 disaster. The daily distance traveled by Narashino residents was collected using geographic information system data from smartphones and processed. These were measured from January 1, 2021, to September 30, 2021. Table 1 presents the mean and standard deviation of the number of people and average travel distance per day by age group. The ages of the residents ranged from teens to seventies, with a mean age and standard deviation of 39.9 ± 20.0 years. The average daily travel distance, representing human flow for Narashino residents, was calculated as the expected value based on the proportion of residents measured per day in each age group and their corresponding travel distances.

2.2. Measures of variables

In this section, the features such as political and legal, as well as health and safety, factors that are expected to have a strong impact on human flow in the COVID-19 disaster are described. The features are presented in Table 2. The categories in Table 2 contain a summary of information and measures reported previously as highly relevant to human flow in the COVID-19 pandemic. In the subcategories (groups) column in Table 2, for each category item, information and measures augmented with the reference date and time, numerical changes, and specific measures for regulating human flow are described. In addition to categories that are likely to be strongly associated with human flow, the breakdown of categories involved further subdivision into several subcategories to elucidate the state of maximum effects. In this study, these groups were subjected to group-level analysis using machine learning to identify groups with particularly strong associations with human flow. The subsubcategory (factors) column in Table 2 provides detailed descriptions of the information and measures, including temporal and spatial data for each group item. These factors were likely to exhibit a high correlation given that the objective was to capture the impact of minor changes in the group items on human flow. Therefore, owing to the potential for unstable training [44] and the bias relating to feature importance [45] in machine-learning models, conducting an appropriate factor analysis is considered challenging. Consequently, the factors were subjected to visual analysis and verified through correlation analysis. The data for this analysis were sourced from the official portals of the Prime Minister's Office of Japan [46], the Ministry of Health, Labour and Welfare of Japan [47], the Tokyo Metropoli-

Table 2. Explanatory variables (features).

Category	Subcategory (groups)[*1]	Subsubcategory (factors)[*2]
Number of newly infected individuals	Number of newly infected individuals on the previous day	Japan Tokyo Chiba
	Difference in the number of newly infected individuals between the previous day and the day before the previous day	Japan Tokyo Chiba
Number of deaths	Number of deaths on the previous day	Japan Tokyo Chiba
	Difference in the number deaths between the previous day and the day before the previous day	Japan Tokyo Chiba
Number of vaccinated individuals	Number of vaccinated individuals in Japan on the previous day	First vaccination Second vaccination
	Cumulative number of vaccinated individuals in Japan up to the previous day	First vaccination Second vaccination
Government infectious disease control measures	Presence/absence of the state of emergency on the day	First implementation in Tokyo Second implementation in Tokyo Third implementation in Tokyo First implementation in Chiba Second implementation in Chiba
	Presence/absence of priority measures to prevent the spread of the disease on the day	First implementation in Tokyo Second implementation in Tokyo First implementation in Chiba
Rate of work-from-home utilization	Rate of work-from-home utilization on the month	Tokyo
Month-end hospital bed occupancy rate	Month-end hospital bed occupancy rate on the month	Japan
Day of the week	Day of the week on the day	Sine wave of the day of the week Cosine wave of the day of the week Weekends & holidays
Climate conditions	Climate conditions on the day	Average temperature in Funabashi Average rainfall in Funabashi Average wind speed in Funabashi

[*1] The day of prediction was used as a reference for the current day, current month, previous day, and the day before the previous day.

[*2] The number of new infections and deaths was segmented by regions for which the information was relevant. The number of vaccinated individuals was segmented by the number of vaccinations. Government infectious disease control measures were segmented according to the region where the measures were implemented and the number of times they were implemented. The day of the week information was segmented by trigonometric values to express periodicity and by weekend and holiday.

tan Government [48], the Japan Meteorological Agency [49], and the Portal Site of Official Statistics of Japan [50] and were processed accordingly. In addition, to evaluate the influence of each feature equally in the analysis, it was necessary to unify the scales of all features. Therefore, standardization with a mean of 0 and a standard deviation of 1 was applied to each feature in this study. The categories and their subdivisions into groups and factors affecting human flow in the COVID-19 pandemic are as follows:

- ***Number of newly infected individuals*** — A close relationship between the number of newly infected individuals and movement of people has been confirmed for the COVID-19 disease. For example, Watanabe and Yabu reported that a 1% increase in the number of newly infected individuals in a prefecture in Japan was associated with a 0.027% decrease in outings in that prefecture [51]. When comparing trends among different regions, incidence rates, which represent the number of infections relative to the population, can be used [52]. However, simply the number of infected individuals is more commonly recognized, because it is encountered by people more often in their daily lives. For this reason, we do not perform transformations such as incidence rates, but adopt these numbers as a single variable. In this study, two variables, the number of newly infected individuals on the previous day and the difference in the number of infected individuals between the previous day and two days before the target prediction day, are used as the group (refer to the "Groups" in Table 2 under "Number of newly infected individuals"). The factors targeted three regions to which the information applies: "Japan," "Tokyo," and "Chiba" (refer to the "Factors" in Table 2 under "Number of newly infected individuals"). The aim was to measure the impact of information on human flow considering infectious diseases in wide and local areas by targeting the entire country, base region for activities other than daily living (region of socioeconomic activities), and base region associated with essential activities of everyday life (region of residence).

- ***Number of deaths*** — Previous studies have reported a close relationship between the number of deaths and human mobility for the COVID-19 disease [53]. When analyzing the impact of fatalities, case fatality rates (CFRs), which represent the number of deaths relative to the number of infected individuals, can be used [54]. However, as with the "Number of newly infected individuals," a simple number of deaths is more commonly recognized, because it is encountered by people more often in their daily lives. For this reason, we do not perform transformations such as CFR, but adopt these numbers as a single variable. In this study, two variables, the number of deaths on the previous day and the difference between the previous day and two days before the target prediction day, are used as the group (refer to the "Groups" in Table 2 under "Number of deaths"). The factors targeted three regions to which the information applies: "Japan," "Tokyo," and "Chiba" (refer to the "Factors" in Table 2 under "Number of deaths"). The aim is the same as that for the "Number of newly infected individuals".

- ***Number of vaccinated individuals*** — Previous studies have reported that vaccination tends to increase human mobility. For example, Liang et al. found that a 10-percentage-point (pp) increase in vaccination coverage in 107 countries was associated with a 1.4–4.3 pp increase in mobility [32]. In this study, two variables, the number of vaccinated individuals on the previous day and the difference between the previous day and two days before the target prediction day, are used as the group (refer to the "Groups" in Table 2 under "Number of vaccinated individuals"). The factors were subdivided based on the number of vaccinations during the measurement period (refer to the "Factors" in Table 2

under "Number of vaccinated individuals").

- ***Government infectious disease control measures*** — Previous studies have reported that public transportation use decreased to less than 20% of the pre-pandemic level due to the state of emergency [55]. In addition, the population density index decreased by 20%, and people tended to avoid moving to densely populated areas or moving between prefectures [56]. In this study, two groups, the state of emergency and priority measures to prevent the spread of disease, were adopted (refer to the "Groups" in Table 2 under "Government infectious disease control measures"). In addition, regarding government infectious disease control measures, it has been noted that the preventive effect varies according to the region of issuance and frequency of implementation [39, 40]. Thus, to consider the individual effects of each factor, each region was further divided according to the implementation times (refer to the "Factors" in Table 2 under "Government infectious disease control measures").

- ***Rate of work-from-home utilization*** — Benita reported that as the number of people opting to work from home (also called telework, remote work, or mobile work) at their discretion or those forced to do so by their companies increases, individual activity and behavior patterns change [57]. In this study, we employed the rate of work-from-home utilization per month in Tokyo to consider the impact of this change on the flow of people (refer to the "Groups" and "Factors" in Table 2 under "Rate of work-from-home utilization").

- ***Month-end hospital bed occupancy rate*** — In Japan, the hospital bed occupancy rate was widely publicized through mass media. Therefore, it is possible that this information may have influenced human mobility. In this study, we used the hospital bed occupancy rate for infectious disease beds in Japan at the end of each month (refer to the "Groups" and "Factors" in Table 2 under "Month-end hospital bed occupancy rate"). When patients are admitted due to infectious diseases (for example, a COVID-19 infection), they are recorded patients in an "infectious disease bed" even when they are admitted to other beds.

- ***Day of the week*** — Irrespective of the COVID-19 pandemic, human migration patterns are strongly influenced by the day of the week [58]. Therefore, we consider this information to stabilize the learning of the prediction models. In this study, the day of the week information of the prediction target day is used as a group (refer to the "Groups" in Table 2 under "Day of the week"), and the day of the week information considering the periodicity, along with the information on weekends and holidays, is used as factors (refer to the "Factors" in Table 2 under "Day of the week"). The day of the week information considering periodicity, which is a factor, was created by performing a feature transformation according to

$$X_{\text{week}}(d) = \cos\left(\frac{2\pi d}{\max(D)}\right), \tag{2.1}$$

$$Y_{\text{week}}(d) = \sin\left(\frac{2\pi d}{\max(D)}\right), \tag{2.2}$$

$$d \in D,\ D = \{0, 1, 2, 3, 4, 5, 6\}, \tag{2.3}$$

[59]. Here, D represents a set of sequence numbers corresponding to the day of the week, with 0 to 6 denoting Monday to Sunday, respectively. Moreover, d is a variable that represents the corresponding

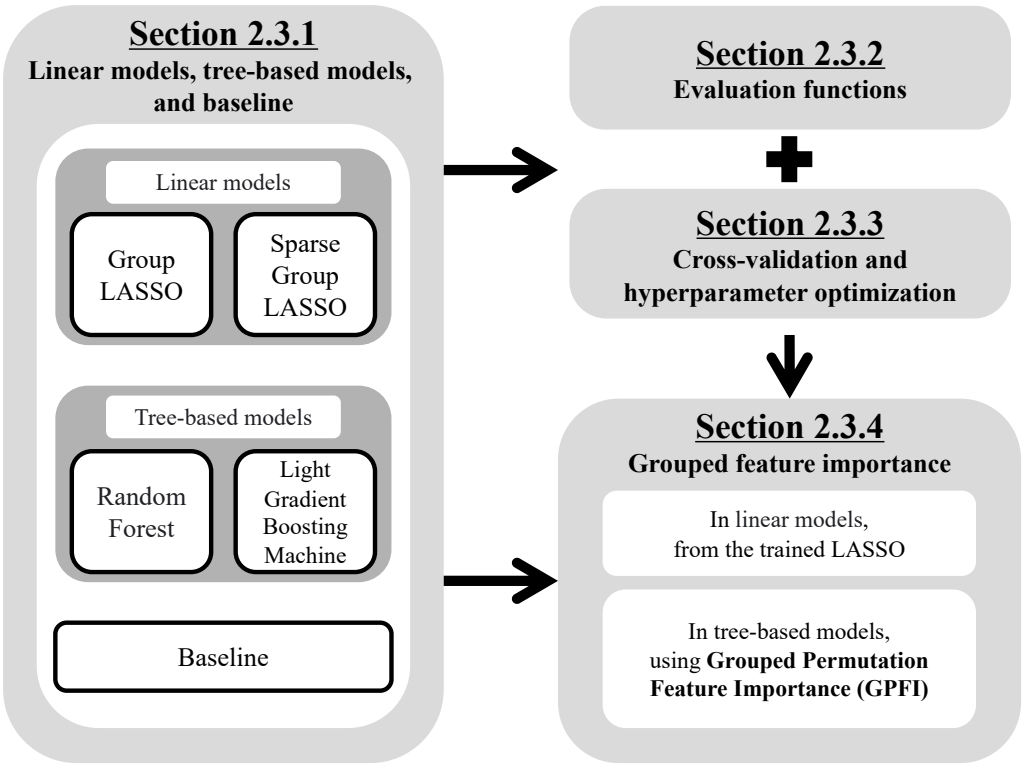


Figure 2. Procedure of the experiment.

day of the week.

- ***Climate conditions*** — Irrespective of the COVID-19 pandemic, it has been reported that temperature, rainfall, and wind speed affect human flow [60]. Therefore, we consider climate conditions to stabilize the learning of the prediction models. Thus, climate information is used as a group (refer to the "Groups" in Table 2 under "Climate conditions"), and temperature, rainfall, and wind speed are used as factors (refer to the "Factors" in Table 2 under "Climate conditions"). For convenience in data preparation, the average daily temperature, rainfall, and wind speed in Funabashi City, Chiba Prefecture, which is adjacent to Narashino City, were used in this study.

2.3. Machine-learning models and data analysis procedures

In this section, an overview of the machine-learning models and data analysis procedures is provided in four subsections. Figure 2 illustrates our experimental procedure. Section 2.3.1 describes the construction of prediction models for human flow using machine learning. In general, the important features and prediction performance differ depending on the machine-learning model used. Therefore, four prediction models were employed in this study: two linear and two tree-based models. A baseline was also employed to confirm whether the models could capture the regularity of the response variable from the features. Section 2.3.2 describes the evaluation functions for a comprehensive comparison of the prediction performance. Section 2.3.3 describes the cross-validation for evaluating the generalization performance of the predictive models and hyperparameter optimization for capturing

the relationship between the features and response variable during training. Section 2.3.4 describes the calculation of grouped feature importance for the linear and tree-based models.

2.3.1. Linear models, tree-based models, and baseline

The least absolute shrinkage and selection operator (LASSO)-based methods, the group LASSO (GL) and the sparse group LASSO (SGL), were selected as the linear models used in this study. GL is a classical LASSO-based method that assigns a penalty to the weights of groups rather than to the individual weights corresponding to each feature [61]. The LASSO regularization term is given by the following equation:

$$\mathrm{LASSO}_{\mathrm{reg}} = \|\boldsymbol{w}\|_1, \;\; \|\boldsymbol{w}\|_1 = \sum_{i=1}^{I} |w_i|, \tag{2.4}$$

where I denotes the total number of features, $\boldsymbol{w}$ denotes the weight vector corresponding to all features used as input, w_i denotes the weight coefficient for the i-th feature, $|\cdot|$ denotes the absolute value, and $\|\cdot\|_1$ denotes the L1-norm. In the LASSO regularization term, the L1 norm, which is defined as the sum of the absolute values of the weight coefficients, is minimized. This approach prevents the divergence of the weight coefficients, enabling the elimination of factors that are unlikely to be strongly associated with the response variable. The GL regularization term is expressed as follows:

$$\mathrm{GL}_{\mathrm{reg}} = \sum_{j=1}^{J} \sqrt{p_j}\, \|\boldsymbol{w}_j\|_2, \;\; \|\boldsymbol{w}_j\|_2 = \sqrt{\sum_{i \in I_j} w_i^2}, \tag{2.5}$$

where J is the total number of groups, $\boldsymbol{w}_j$ is the weight vector for the features in the j-th group, I_j is the set of features in the j-th group, w_i is the weight coefficient for the i-th feature in the I_j group, and $\|\cdot\|_2$ is the L2-norm. p_j is the size of the j-th group, and multiplying it by $\sqrt{p_j}$ accounts for the variability in each group size. In the GL regularization term, the L2-norm is calculated for each group and the sum of these norms is minimized. This approach facilitates the elimination of groups that are unlikely to be strongly associated with the response variable. GL was implemented using the Python package "group-lasso" (version 1.5.0). SGL is a LASSO-based method that allows not only the assignment of a penalty to each group but also the assignment of a penalty to each individual weight [62, 63]. The SGL regularization term is expressed as follows:

$$\mathrm{SGL}_{\mathrm{reg}} = \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \sum_{j=1}^{J} \sqrt{p_j}\, \|\boldsymbol{w}_j\|_2, \tag{2.6}$$

where the coefficients of λ_1 and λ_2 are applied as multipliers to the respective regularization terms for LASSO and GL and then summed. This enables the calculation of feature importance for each feature in addition to the calculation of the importance of each group. SGL was implemented using the Python package "group-lasso" (version 1.5.0).

Random forest (RF) and light gradient boosting machine (LightGBM) were selected as the tree-based models in this study. RF extends the bagging method, which is a type of ensemble learning method [64] consisting of multiple trees (decision trees for classification and regression trees for regression). The training of each tree utilizes subsets of data extracted from the dataset using the bootstrap method. In addition, for each branch in the tree, only a certain number of features that are selected

randomly, rather than the entire set, are used in training. The final output value is derived from the leaf node of each trained tree, determined through majority voting for classification or by calculating the mean of all outputs for regression. Introducing randomness in this manner reduces the correlation among trees, thereby enhancing the bias and variance relative to using a singular decision tree. RF was implemented using the Python package "scikit-learn" (version 1.0.2) [65]. LightGBM is a type of gradient boosting decision tree that addresses the computational cost issues associated with traditional models of this type [66]. Similar to RF, it comprises multiple trees for both classification and regression. However, unlike RF, which predicts the actual value of the response variable, LightGBM focuses on predicting the residual between the actual target and output values. The trees are combined in series, allowing subsequent trees to correct the data samples with large residuals identified by the preceding tree preferentially. By leveraging the training outcomes from one tree to the next, both the training efficiency and prediction performance are enhanced. LightGBM incorporates several computational performance improvement techniques, including leaf-wise tree growth, histogram- and gradient-based one-side sampling, and exclusive feature bundling (detailed information can be found in the original paper [66]). This model was implemented using the Python package "lightgbm" (version 4.0.0).

In this study, the baseline was established as the mean of the response variable to simulate the prediction performance when the association between the feature and response variable was not captured. Comparing the outcomes of the four machine-learning models with the baseline facilitated the assessment of the prediction performance and the degree to which the features that were associated with the response variable were captured. If the prediction performance of the machine-learning model was better than that of the baseline, it could be inferred that the model successfully captured the association between the response variable and features to some extent.

2.3.2. Evaluation functions

In this study, the root mean square error (RMSE), coefficient of determination (R-squared: R^2), and mean absolute percentage error (MAPE) were employed as evaluation functions to compare the prediction performance of the machine-learning models comprehensively.

RMSE is an evaluation function expressed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{2.7}$$

where N is the sample size, y_i is the measured value for the i-th data point, and $\hat{y}_i$ is the predicted value for the i-th data point. The advantage is that the least-squares error between the measured and predicted values is multiplied by the square root. Therefore, the units of the evaluation value can be evaluated on the same scale as the units of the actual data. It is possible to determine that the closer the evaluation value is to 0, the better the prediction performance.

R^2 is an evaluation function expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2.8}$$

where N is the sample size, y_i is the measured value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, and $\bar{y}$ is the average of the measured values. The denominator of the second term,

$\sum_{i=1}^{N}(y_i-\bar{y})^2$, represents the sum of the variations from the mean in the measured values. The numerator of the second term, $\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$, represents the sum of the residuals, where the measured values cannot be explained by the predicted values (or the machine-learning model and features). That is, the second term indicates the ratio of the sum of the residuals that cannot be explained by the predicted values to the sum of the variations from the mean in the measured values. Therefore, R^2, where the second term is subtracted from 1, is interpreted as the ratio of the predicted values to the variation from the mean in the measured values. The maximum value is 1, and it is possible to determine that the closer the evaluation value is to 1, the better the prediction performance.

MAPE is an evaluation function expressed as

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right|, \tag{2.9}$$

where N is the sample size, y_i is the measured value for the i-th data point, and $\hat{y}_i$ is the predicted value for the i-th data point. The advantage is that the prediction error divided by the measured value is multiplied by the absolute value and 100. Therefore, the prediction error relative to the measured value can be evaluated as a positive percentage. Because a smaller error percentage is desirable, the closer it is to zero, the better the prediction performance.

2.3.3. Cross-validation and hyperparameter optimization

In this study, K-fold cross-validation (CV) was applied when training the four machine-learning models. Two types of CV were performed simultaneously: CV on the training and test data, and CV on the training and validation data. The CV on the training and test data was performed to evaluate the generalization performance of the prediction models. The CV on the training and validation data was used to perform hyperparameter optimization based on the validation error minimization, as described below. In this case, the dataset was first randomly shuffled to eliminate bias. Subsequently, $K = 10$ was employed as the number of dataset partitions, and the original dataset was divided in a training:test ratio of 9:1. In addition, the training data were re-split in a training:validation ratio of 9:1. Therefore, when CV on the training and test data was run once, one prediction model was built for the training and test data, whereas 10 models were built for the validation data. Because the CV on the training and test data was run 10 times, the final evaluation value was the output from 10 prediction models for the training and test data, and 100 for the validation data. The generalization performance of the prediction models constructed using the given hyperparameters and different datasets was evaluated comprehensively by calculating the summary statistics of these output values.

Hyperparameter optimization based on minimizing the validation error was performed to enhance the ability of the machine-learning models to capture the features of the dataset. Bayesian optimization was used as the optimization algorithm and RMSE was used as the optimization metric. In GL, the regularization parameter, which determines the strength of the regularization term, was targeted for optimization, with the maximum number of training iterations set to 100,000. In SGL, the two regularization parameters that determine the strength of the LASSO and GL regularization terms were targeted for optimization, with a maximum of 80,000 training iterations. In RF, the optimization covered six parameters: the number of trees to be constructed, data splitting criteria, maximum depth of the tree, minimum sample size for a split, minimum sample size for a leaf node, and penalty for tree

complexity. In LightGBM, the nine parameters targeted for optimization were the number of trees to be constructed, final number of leaf nodes, maximum tree depth, minimum sample size for a leaf node, L1 regularization coefficient, L2 regularization coefficient, fraction of data randomly sampled for tree construction, frequency of bagging, and fraction of features used for tree construction. These hyperparameters were optimized using Optuna [67]. For other hyperparameters, the default values of the Python package for each of the machine-learning models described in Section 2.3.1 were used.

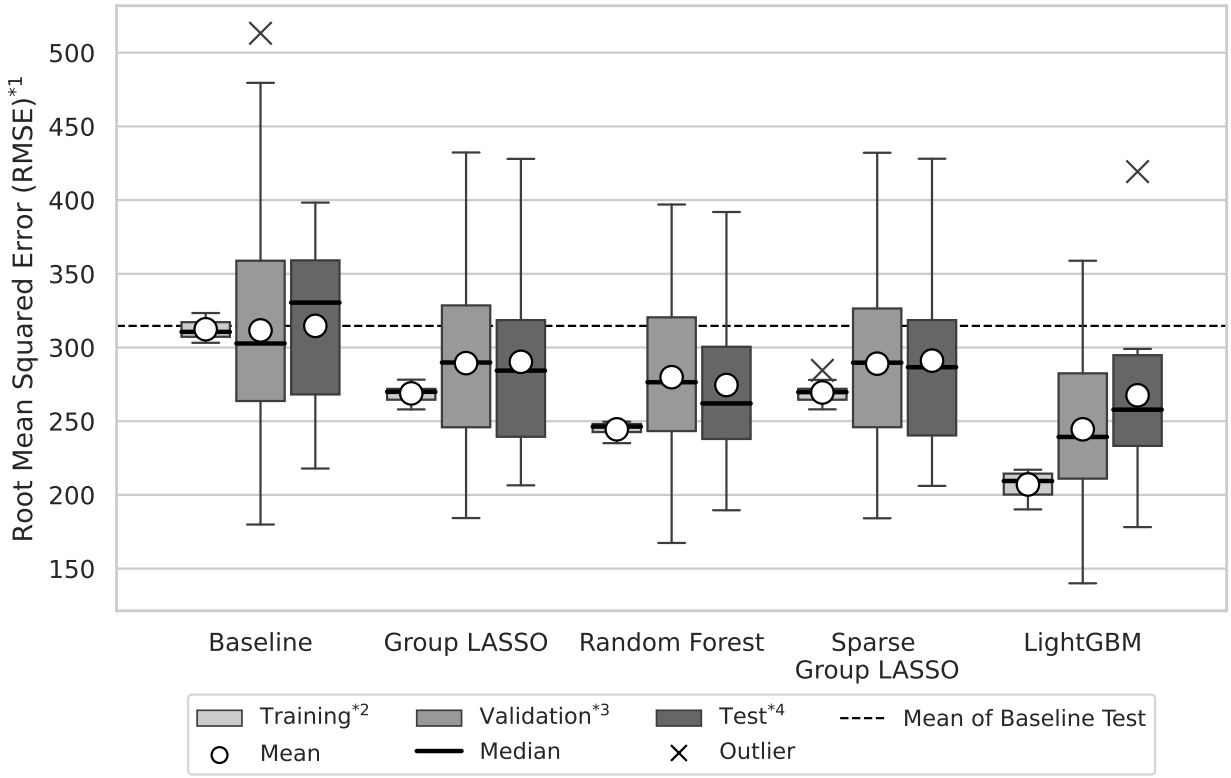
2.3.4. Grouped feature importance

This section describes the evaluation method for the grouped feature importance using a trained machine-learning model. Extensions of the LASSO model were used as linear machine-learning models. It was therefore possible to evaluate the feature importance of the groups by referring to the weight coefficients obtained after training the model. For tree-based models, although the importance of each feature can be calculated separately, the feature importance of a group cannot be determined. Therefore, the grouped permutation feature importance (GPFI) [68] was adopted for the tree-based models in this study. GPFI calculates the feature importance for groups by shuffling feature values in group units and evaluating them using a trained machine-learning model. In particular, the pre- and post-shuffle datasets are used as input into a trained machine-learning model, and the evaluation values for each are computed. Subsequently, the difference between the two evaluation values is computed and set as the feature importance for that group. The larger the difference between the two values, the greater the dependence on the group, enabling it to be determined as an important group. Conversely, the smaller the difference between the two values, the lower the dependence on the group, making it unimportant.

The feature importance for groups is likely to vary depending on the training data. In this study, the feature importance for groups was computed in each CV iteration. By computing the summary statistics of these values, the variability in the feature importance depending on the dataset was considered. Moreover, to enable an intuitive judgment of feature importance, the feature importance for the groups was normalized to have a maximum value of 1 and minimum value of 0.

3. Results

- We confirm that all prediction models were properly trained. For this purpose, we assessed the possibility of overfitting by referencing the RMSE, which was used as an optimization index for hyperparameter optimization. Boxplots of the RMSE for the training, validation, and test data for each prediction model are shown in Figure 3. In general, for a given sample size, employing high-capacity machine-learning models and/or data featuring high dimensionality can lead to overfitting, thereby impairing the ability of the model to generalize to new, unseen data effectively. In such a case, there may be a large discrepancy between the evaluation values of the training and validation data or between the validation and test data. In addition, the evaluation value of the validation data is much worse than that of the training data, and the evaluation value of the test data is much worse than that of the validation data. Based on the above, with a focus on GL and SGL, which are the linear models in Figure 3, it was confirmed that the interquartile ranges of the evaluation values in the training, validation, and test data almost overlapped. Thus, it is considered that the linear models were not likely to exhibit overfitting and that they were properly trained. Next, with a focus on RF and LightGBM, which are the tree-based models shown in Figure 3, it was confirmed



*1 An RMSE close to 0 indicates better prediction performance.
*2 Training shows the boxplots of RMSE for the training data used in the 10 prediction models constructed by K-fold CV in training/test with respect to each machine-learning model.
*3 Validation shows the boxplots of RMSE for the validation data used in the 100 prediction models constructed by K-fold CV in training/test and training/validation with respect to each machine-learning model.
*4 Test shows the boxplots of RMSE for the test data used in the 10 prediction models constructed by K-fold CV in training/test with respect to each machine-learning model.

Figure 3. Boxplots of RMSE: the evaluation value of each machine-learning model obtained from K-fold CV.

that the interquartile ranges of the training and validation data partially overlapped. In addition, the interquartile ranges of the evaluation values in the validation and test data overlapped widely. Therefore, it can be concluded that there was a slight tendency for the tree-based models to fit the training data; however, this tendency is considered minor. Furthermore, the wide overlap between the interquartile ranges of the validation and test data suggests that the tree-based models exhibited good generalization performance for unknown data. These results suggest that the tree-based models were unlikely to suffer from overfitting, indicating that the training was conducted appropriately. Following this, the mean (median) ± standard deviation of RMSE, R^2, and MAPE for the baseline, GL, RF, SGL, and LightGBM models are shown in Table 3. First, we focus on all evaluation values for the baseline and each prediction model in Table 3. The mean and median of RMSE, R^2, and MAPE for all prediction models were superior to those of the baseline, regardless of the training, validation, or test data. This suggests that the linear and tree-based models could capture patterns that explain human flow from the given features. Next, we focus on the GL, RF, SGL, and LightGBM

Table 3. Mean (median) ± SD of prediction scores for each machine-learning model.

Evaluation function[*1]	Prediction model	Mean (median) ± SD		
		Training[*2]	Validation[*3]	Test[*4]
RMSE	Baseline	312 (311) ± 6.7	312 (303) ± 68.6	315 (330) ± 59.0
	GL	269 (270) ± 6.3	289 (290) ± 55.3	290 (284) ± 63.3
	RF	244 (246) ± 5.0	280 (276) ± 48.0	275 (262) ± 54.5
	SGL	269 (270) ± 7.4	289 (290) ± 55.7	291 (287) ± 63.0
	LightGBM	**207 (209) ± 8.7**	**244 (239) ± 45.8**	**268 (258) ± 62.7**
R^2	Baseline	.000 (.000) ± .000	.000 (.000) ± .000	.000 (.000) ± .000
	GL	.620 (.619) ± .014	.488 (.526) ± .163	.496 (.523) ± .212
	RF	.686 (.683) ± .010	.503 (.585) ± .218	.553 (.599) ± .186
	SGL	.618 (.619) ± .016	.489 (.526) ± .164	.494 (.517) ± .212
	LightGBM	**.774 (.771) ± .020**	**.626 (.683) ± .150**	**.565 (.628) ± .207**
MAPE	Baseline	.057 (.057) ± .001	.057 (.056) ± .014	.057 (.058) ± .012
	GL	.036 (.036) ± .001	.040 (.039) ± .007	.040 (.038) ± .008
	RF	.033 (.033) ± .001	.038 (.038) ± .007	.038 (.035) ± .008
	SGL	.036 (.036) ± .001	.040 (.039) ± .007	.040 (.039) ± .008
	LightGBM	**.028 (.028) ± .001**	**.034 (.033) ± .006**	**.037 (.036) ± .008**

[*1] An RMSE close to 0 indicates better prediction performance. An R^2 close to 1 indicates better prediction performance. A MAPE close to 0 indicates better prediction performance.

[*2] The Training column shows the summary statistics of RMSE, R^2, and MAPE for the training data used in the 10 prediction models constructed by K-fold CV in training/test with respect to each machine-learning model.

[*3] The Validation column shows the summary statistics of RMSE, R^2, and MAPE for the validation data used in the 100 prediction models constructed by K-fold CV in training/test and training/validation with respect to each machine-learning model.

[*4] The Test column shows the summary statistics of RMSE, R^2, and MAPE for the test data used in the 10 prediction models constructed by K-fold CV in training/test with respect to each machine-learning model.

models in Table 3. In the training, validation, and test data, the mean and median values of RMSE, R^2, and MAPE of the tree-based models were superior to those of the linear models. Among the tree-based models, LightGBM had the best mean and median RMSE, R^2, and MAPE for the training, validation, and test data.

- Among the four machine-learning models, LightGBM exhibited the best prediction performance. This suggests that LightGBM is superior to other machine-learning models in capturing critical information relating to the response variable from features. Accordingly, factor analysis was conducted to examine the impact on human flow for each group based on the group feature importance of LightGBM. The summary statistics of the group feature importance for each iteration of the K-fold CV are shown in Table 4. Normalization within the range of [0, 1] was applied to each iteration to highlight the differences in feature importance. In this study, the open interval between A and B is denoted by (A, B) and the closed interval is denoted by $[A, B]$. First, focusing on the mean values of the feature importance displayed in Table 4, it is observed that for all groups the mean was greater than 0. When the importance of a feature that has no impact on the prediction was set to 0, the normalized values for it resulted in both a mean and standard deviation of 0.000. These results suggest that all groups had a degree of association with human flow during the COVID-19 epidemic, albeit small. However, for state-of-emergency and priority measures to prevent the spread of disease, as shown in Table 4, the mean was nearly zero. Next, focusing on ranking based on the mean values of the group feature importance in Table 4, we identified the cumulative number of vaccinated individuals as the primary contributor to human flow. The climate conditions were selected as the second

Table 4. Summary statistics for feature importance at the group level with [0, 1] normalization applied to each iteration of cross-validation and group rankings based on the mean value of feature importance.

Groups[*1]	Mean[*2]	SD[*2]	Ranking[*3]
Number of newly infected individuals on the previous day	.257	.144	4
Difference in the number of newly infected individuals between the previous day and the day before the previous day	.021	.018	9
Number of deaths on the previous day	.087	.056	6
Difference in the number deaths between the previous day and the day before the previous day	.022	.016	8
Number of vaccinated individuals in Japan on the previous day	.354	.330	3
Cumulative number of vaccinated individuals in Japan up to the previous day	.929	.226	1
Presence/absence of the state of emergency on the day	.002	.002	11
Presence/absence of priority measures to prevent the spread of disease on the day	.002	.003	11
Rate of work-from-home utilization on the month	.025	.035	7
Month-end hospital bed occupancy rate on the month	.004	.006	10
Day of the week on the day	.252	.089	5
Climate conditions on the day	.586	.291	2

[*1] The day of prediction was used as a reference for the current day, current month, previous day, and the day before the previous day.
[*2] When the feature importance was set to 0 to imply no effect on prediction, applying normalization, both the mean and standard deviation were 0.000.
[*3] Ranking were based on the mean value of feature importance.

most important group, followed by the number of vaccinated individuals, number of newly infected individuals, and day of the week. Excluding the climate conditions and day of the week, which were added to stabilize the learning of each prediction model, groups with strong associations to human flow during the COVID-19 pandemic were the cumulative number of vaccinated individuals, number of vaccinated individuals, and number of newly infected individuals. However, Table 4 shows that the number of vaccinated individuals had the highest standard deviation among all groups, indicating that the stability of this group was low. Conversely, the number of newly infected individuals and cumulative number of vaccinated individuals had small standard deviations from the mean, suggesting that they could stably capture the trends in human flow. Therefore, it is highly likely that the number of newly infected individuals and cumulative number of vaccinated individuals had a strong and stable influence on human flow during the COVID-19 pandemic.

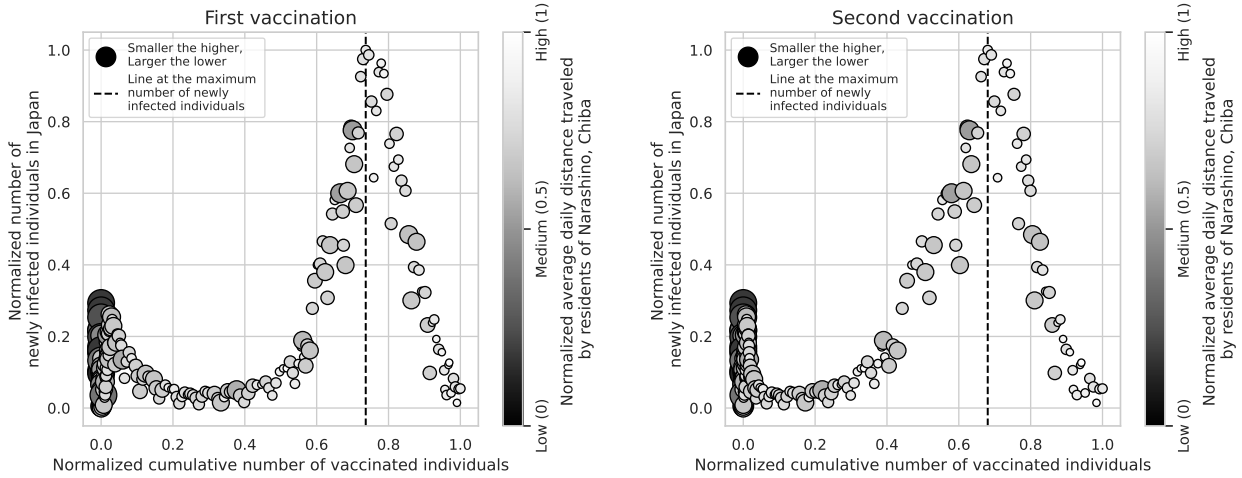


Figure 4. Scatter plots of the normalized number of newly infected individuals in Japan, normalized cumulative number of vaccinated individuals broken down by the number of vaccinations, and normalized average daily distance traveled.

- The number of newly infected individuals and cumulative number of vaccinated individuals had a strong and stable effect on human flow in the COVID-19 disaster. A visual analysis using scatter plots is presented to capture these detailed trends, which are subdivided into the factors listed in Table 2. Figure 4 shows scatter plots of the number of newly infected individuals by region (in this figure, Japan), cumulative number of vaccinated individuals subdivided by the number of vaccinations, and human flow of Narashino residents. For uniformity, all values were normalized using [0, 1] scaling. Focusing on Figure 4, it was confirmed that the human flow tended to be low when the normalized cumulative number of vaccinated individuals was 0, regardless of the number of vaccinations. Conversely, when the normalized cumulative number of vaccinated individuals was greater than 0, the human flow also tended to increase. In addition, there may be a difference in the human flow trend in relation to the increase or decrease in the normalized number of newly infected individuals before and after the maximum normalized number of newly infected individuals (see the dashed line in Figure 4). In particular, the human flow tended to increase after the peak compared to before the peak, as the normalized number of newly infected individuals decreased. These trends are also confirmed by scatter plots consisting of the normalized number of newly infected individuals in the Tokyo and Chiba Prefectures (see Figure A2 in the Appendix). To analyze these trends in further detail, the segments were divided according to the cumulative number of vaccinated individuals. The correlation between the number of newly infected individuals and human flow in each segment was then confirmed. The three segments are as follows: the normalized cumulative number of vaccinated individuals = 0 (segment A); the normalized cumulative number of vaccinated individuals > 0 but below the threshold value (segment B); and the normalized cumulative number of vaccinated individuals > the threshold value but below 1 (segment C). The threshold value refers to the cumulative number of vaccinated individuals when the normalized number of newly infected individuals is at its maximum in the respective scatter plots. For example, the normalized cumulative number of vaccinated individuals at the first vaccination dose was 0.737 and the normalized cumulative number of vaccinated individuals at the second vaccination dose was 0.681 when the number of

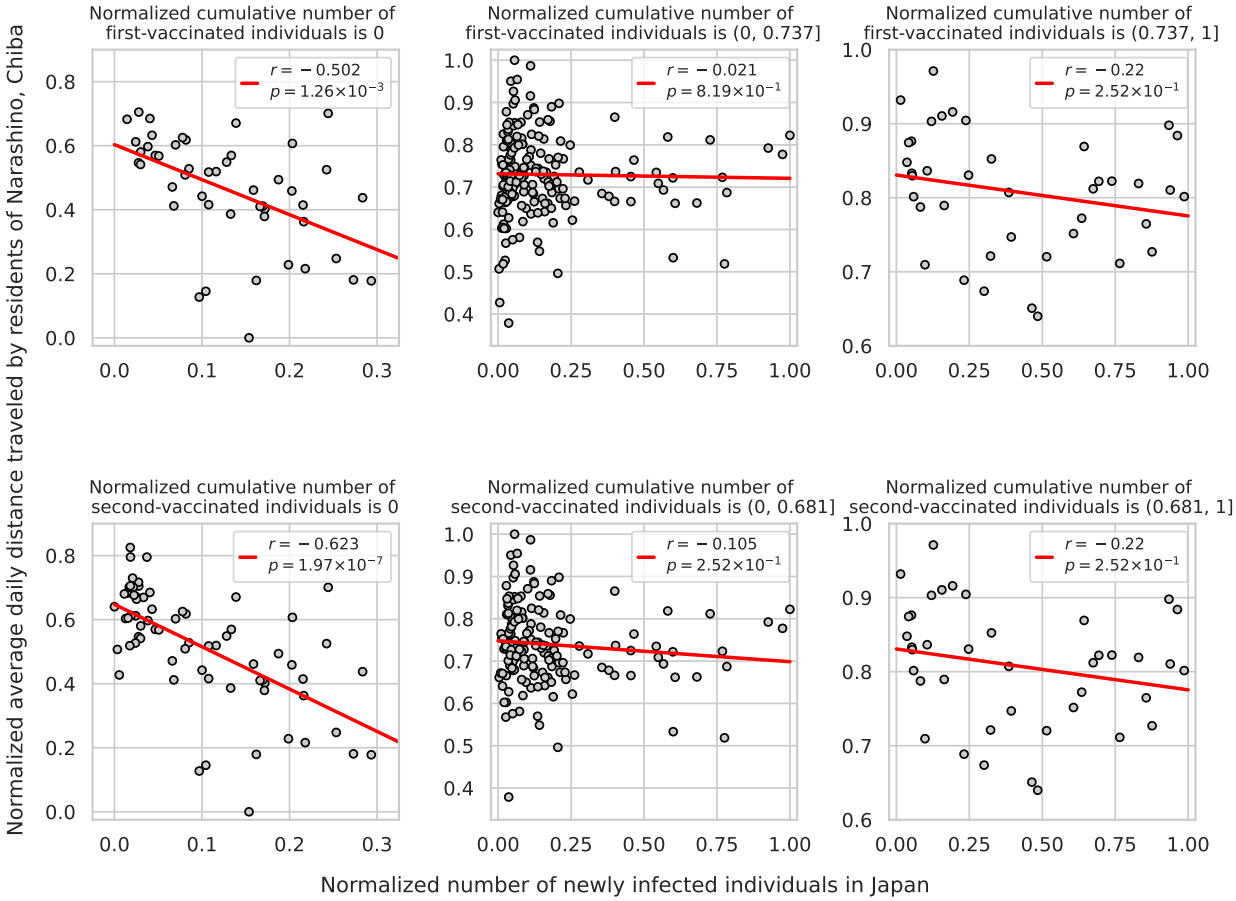
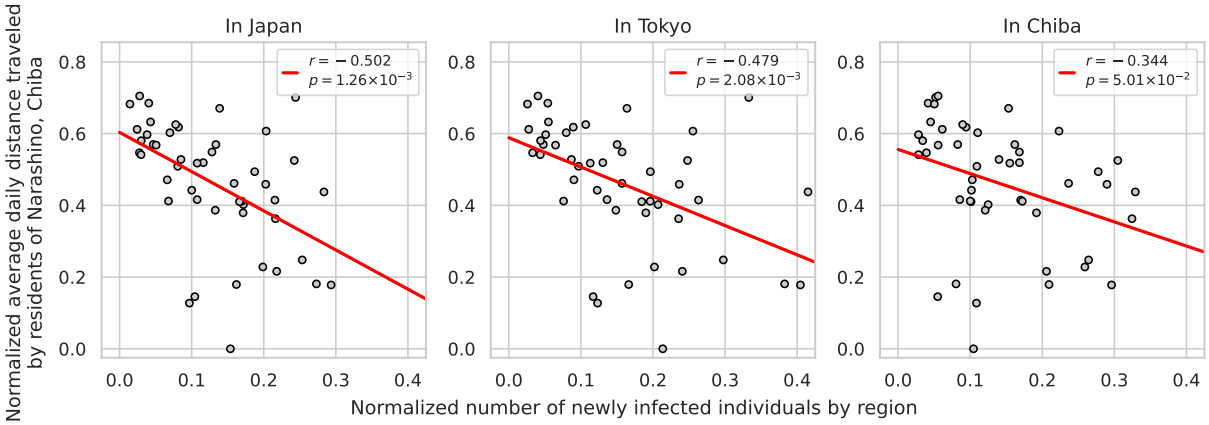
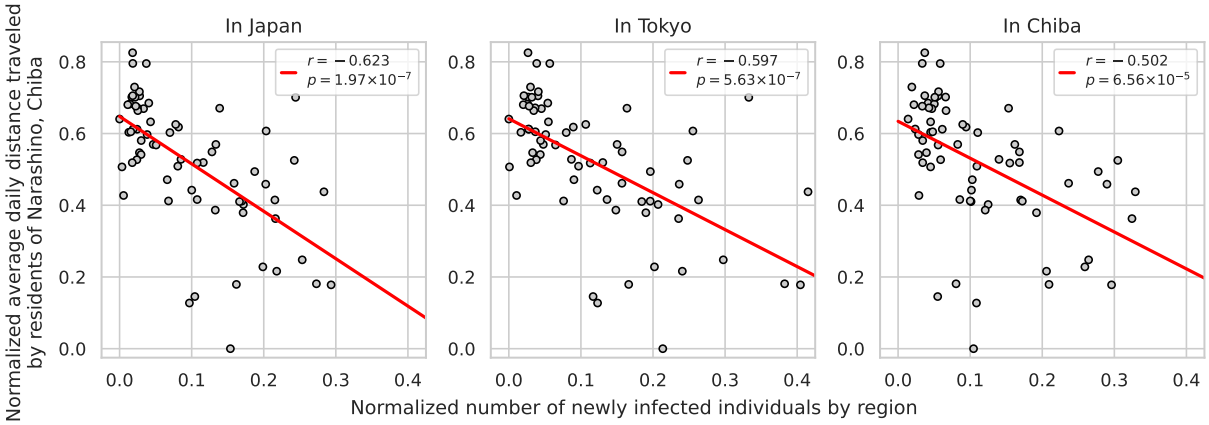


Figure 5. Scatter plots, correlations, and adjusted p-values of the normalized number of the newly infected individuals in Japan, normalized average daily distance traveled broken down by the number of vaccinations, and normalized cumulative number of vaccinated individuals by segment (left: segment A, middle: segment B, right: segment C).

newly infected individuals in Japan reached its maximum (see the dashed line in Figure 4). Scatter plots of the number of newly infected individuals in Japan and human flow, divided into segments according to the above, and their correlation coefficients r are presented in Figure 5 (those consisting of the number of newly infected individuals in Tokyo and Chiba Prefecture are shown in Figure A2 in the Appendix). All values were normalized to a scale of [0, 1] to ensure uniformity. In addition, to determine the significance of the correlation coefficients in this study, a test of no correlation was conducted with the null hypothesis being "correlation coefficient is 0" and the alternative hypothesis being "correlation coefficient is not 0." The significance level was set to $\alpha = 0.05$. Furthermore, the Benjamini–Hochberg (BH) method, which is a multiple testing correction, was applied to address the problem of multiplicity, in which the probability of an error of the first kind (false positive) increases with repeated multiple times. The p-values adjusted using the BH method are shown in Figure 5 (the adjusted p-values for Tokyo and Chiba are shown in Figure A2 in the Appendix). We focus on the correlation coefficients r and adjusted p-values for each segment in Figure 5. For the first vaccination, in the upper left of segment A in Figure 5, the correlation coefficient was −0.502,



(a) First vaccination



(b) Second vaccination

Figure 6. Scatter plots, correlations, and adjusted p-values between the normalized number of newly infected individuals in Japan (left), Tokyo (center), and Chiba (right) and the normalized average daily distance traveled when the cumulative number of vaccinated individuals was 0.

with an adjusted p-value of 1.26×10^{-3}. For the second vaccination, in the lower left of segment A in Figure 5, the correlation coefficient was -0.623 and the adjusted p-value was 1.97×10^{-7}. Therefore, the number of newly infected individuals and human flow may be negatively correlated in segment A, rather than uncorrelated. In contrast, in segment B, in the center of Figure 5, the correlation coefficient was -0.021 for the first vaccination and -0.105 for the second vaccination, with adjusted p-values of 0.819 and 0.252, respectively. Thus, the possibility that the relationship between the number of newly infected individuals and human flow was independent was confirmed. Segment C, on the right side of Figure 5, had a correlation coefficient of -0.220, showing a stronger negative correlation than segment B. However, the adjusted p-value of 0.252 confirms the possibility that the relationship between the number of newly infected individuals and human flow was independent. Similar trends were also observed in Tokyo and Chiba. However, for the first vaccination of segment A in Chiba, the adjusted p-value was 5.01×10^{-2}, confirming that the relationship between the number of newly infected individuals and human flow was statistically uncorrelated (Figure A2(b),

upper left in the Appendix).

- We investigate how differences in the number of newly infected individuals by region and the number of vaccinations influence human flow. Figure 6 shows scatter plots of the number of newly infected individuals by region and human flow, as well as the correlation coefficients, under the situation in segment A where vaccination had not started. First, we focus on the difference in the number of vaccinations. Comparing Figure 6(a),(b), it is confirmed that the correlation coefficient was stronger in the negative direction for the second vaccination than for the first in all regions. We examined whether there was a significant difference in the correlation coefficients owing to differences in the number of vaccinations. For each region, a test of the difference between independent correlation coefficients based on the z-distribution was conducted, where the null hypothesis was "no difference in correlation coefficients" and the alternative hypothesis was "there is a difference in correlation coefficients." The significance level was set to $\alpha = 0.05$. Furthermore, the BH method was applied to address the multiplicity problem. The adjusted p-values were 0.194 for Japan, 0.194 for Tokyo, and 0.194 for Chiba. Therefore, there was no significant difference in the correlation coefficients between the number of newly infected individuals and human flow according to the number of vaccinations in all regions. Next, we focus on the differences in the number of newly infected individuals by region. Regarding the first vaccination in Figure 6(a), it was confirmed that the correlation coefficient between the human flow in Narashino City and number of newly infected individuals in Japan was -0.502, that in Tokyo was -0.479, and that in Chiba was -0.344. Focusing on the second vaccination in Figure 6(b), it was confirmed that the correlation coefficient between the human flow in Narashino City and number of newly infected individuals in Japan was -0.623, that in Tokyo was -0.597, and that in Chiba was -0.502. Therefore, in the situation where vaccination had not yet begun, the negative correlation with human flow was greater for the number of newly infected individuals in major regions such as Tokyo and Japan than in Chiba, where Narashino residents live. It was necessary to confirm whether there were significant differences in the correlation coefficients between regions. For each number of vaccinations, a test of the difference between the independent correlation coefficients based on the z-distribution was conducted, where the null hypothesis was "no difference in correlation coefficients" and the alternative hypothesis was "there is a difference in correlation coefficients." The significance level was set to $\alpha = 0.05$. Furthermore, the BH method was applied to address the multiplicity problem. The adjusted p-values for the first vaccination were 0.443 for Japan and Tokyo, 0.330 for Japan and Chiba, and 0.330 for Tokyo and Chiba. The adjusted p-values for the second vaccination were 0.443 for Japan and Tokyo, 0.330 for Japan and Chiba, and 0.330 for Tokyo and Chiba. Thus, we confirmed that there were no significant differences between the correlation coefficients of the number of newly infected individuals by region and human flow, regardless of the number of vaccinations.

4. Discussion

4.1. Explanation of results

Table 5 shows the results obtained from this study.

(i). All machine-learning models were properly trained with a low probability of overfitting. In addition, it was confirmed that the tree-based models scored better than the linear models for all

Table 5. Summary of the results of this study.

(i).	All predictive models were found to learn adequately, with negligible possibility of over-learning. It was also confirmed that LightGBM had the best prediction performance in terms of all evaluation functions, RMSE, R^2, and MAPE.
(ii).	The groups in Table 2 influence human flow in the COVID-19 pandemic, but the contribution of the government infectious disease control measures was confirmed to be relatively low. The results also suggest that the cumulative number of vaccinated individuals and number of newly infected individuals may have a stable and strong influence on the increase or decrease in human flow.
(iii).	Before and after the cumulative number of vaccinated individuals is zero, the trends in human flow differ. Similarly, these trends also differ before and after the peak in the number of newly infected individuals. The statistical results show that there is a correlation between the number of newly infected individuals and human flow in each region when the first and second vaccinations had not started ($p < 0.05$) (however, it was suggested that there may be no correlation between the number of newly infected individuals and human flow in Chiba when the first vaccination had not started ($p \geqq 0.05$)).
(iv).	The negative correlation coefficients between the number of newly infected individuals and human flow in each region were larger for the second vaccination than that for the first vaccination under conditions in which vaccination had not started (however, the difference in the correlation coefficients for the vaccination frequency was not significant ($p \geqq 0.05$)).
(v).	The negative correlation between the number of newly infected individuals and human flow was higher in major regions such as Tokyo and Japan than in Chiba, where people live (however, there was no significant difference between the correlation coefficients of the number of newly infected persons and human flow by region ($p \geqq 0.05$)).

three evaluation functions (RMSE, R^2, and MAPE) among the five prediction models, including the baseline. This may be owing to the ability of tree-based models to capture the interactions among input variables and the nonlinear relationships between the features and response variable. Tree-based models can effectively learn how other features influence the response variable when a feature has a particular value. Even when the influence involves nonlinearity, the pattern of influence can be learned. The features in this study had complex interactions, and there were likely to be nonlinear relationships between the features and response variable. However, the linear models employed in this study were unable to capture the interactions between the input variables and nonlinear relationships effectively. This may explain why the prediction performance of the tree-based models was better than that of the linear models. Among the tree-based models, LightGBM was found to have the best prediction performance. The superior prediction performance of LightGBM over RF is likely attributable to the differences in the model characteristics. The prediction errors of machine-learning models have biases and variances. Bias

represents the difference between the true and predicted values. The implication is that the lower the bias, the better the capturing of the association between the response variable and features of a given dataset. Variance represents the scatter in the predicted values. The implication is that the lower the variance, the better the capturing of the association between the response variable and features common to a wide range of data, not only a specific dataset. In general, there is a trade-off between bias and variance, and it is crucial to find models in which both parameters are small [69]. RF combines multiple independent weak learners in parallel, where their predicted values undergo majority voting (for classification) or averaging (for regression) to determine the final outcome. This approach implies high generalization performance as it integrates the collective decisions of weak learners, thereby making RF a model that excels in reducing variance rather than bias. In contrast, LightGBM constructs decision trees with dependent relationships in series, training by adjusting the sample based on residuals by the preceding decision tree in the subsequent tree. This method, which focuses on individual samples, suggests that LightGBM is adept at reducing bias rather than variance. Therefore, although LightGBM tended to have a higher standard deviation of prediction scores than RF, its lower mean indicates better performance. Previous studies have reported that LightGBM outperforms RF in regression prediction in various fields [70–72].

(ii). The importance of group features, focusing on LightGBM, which demonstrated the best prediction performance, was confirmed. Table 2 shows that all groups had an impact on human flow in the event of a COVID-19 disaster. However, government infectious disease control measures contributed little to the prediction of human flow. This outcome is likely owing to a diminishing response to each intervention [73], changes in human flow resulting from alterations in requests and orders [74], and an increase in risk-compensatory behavior following the enforcement of regulations [75]. The results of this study support these claims and further show the possibility of their significant influence. By contrast, the cumulative number of vaccinated individuals and number of newly infected individuals were found to have a potentially stable and strong impact on human flow. The Peltzman effect, which refers to the phenomenon in which safety measures decrease an individual's perception of risk, leading to potentially riskier decisions [76], is a probable explanation for the emphasis on the numbers of vaccinated and newly infected individuals. Previous studies reported an increase in the Peltzman effect with vaccinations during the COVID-19 outbreak [32,77,78]. The existence of a stochastic trend and periodicity has also been reported in which the number of newly infected individuals increases as human mobility increases, while people tend to curb their movements in response to an increase in newly infected individuals [79]. Therefore, it is possible that the Peltzman effect of the COVID-19 pandemic was influenced by a decrease in the understanding of infection risk owing to an increase in the cumulative number of vaccinated individuals and a change in the understanding of infection risk owing to cycles of increase and decrease in the number of newly infected individuals. These results are consistent with those reported in previous studies, and are considered highly valid.

(iii). The visual analysis qualitatively confirmed that there was a difference in the trend of human flow between the times when the cumulative number of vaccinated individuals was 0 and when it was higher than 0, as well as before and after reaching the maximum number of newly infected individuals. The correlation analysis quantitatively investigated this trend and found that

the number of newly infected individuals and human flow were not "no correlation" when the cumulative number of vaccinated individuals was 0 (segment A), indicating the possibility of a negative correlation. In addition, we confirmed that the relationship between the number of newly infected individuals and human flow can indicate no correlation in the range where the cumulative number of vaccinated individuals was greater than 0 and the number of newly infected individuals reached the maximum (segment B). Furthermore, in the range beyond that (segment C), a stronger negative correlation between the number of newly infected individuals and human flow was observed compared with that in segment B. However, the correlation between the number of newly infected individuals and human flow was found to be insignificant. This suggests that the Peltzman effect in the COVID-19 pandemic may be closely related to human flow. In segment A, the cumulative number of vaccinated individuals was 0, which means that the perception of infection risk owing to an increase in the number of new infections was high. In segment B, the increase in the cumulative number of vaccinated individuals was considered to reduce the perception of infection risk owing to the increase in the number of newly infected individuals. In segment C, an infectious explosion was experienced. However, the further increase in the cumulative number of vaccinated individuals may have prevented an increase in the perception of infection risk owing to the number of newly infected individuals. These results seem to support the claim of a Peltzman effect induced by vaccination and the number of newly infected individuals. In segment A, people were aware of the risk of infection, and movement restrictions appeared to be working effectively. In contrast, vaccination had commenced during segments B and C. However, people may have been less conscious of the risk of infection, as they determined their daily travel distance, regardless of the number of newly infected individuals. Kwon and Koylu reported that the correlation between COVID-19 infection rates and human mobility weakens over time [80]. In addition, they attributed this to changes in disease control measures, risk perception, and individual behavior, which vary spatially and temporally. Serisier et al. reported an increased tendency to interact with individuals outside the home and to utilize non-essential stores and services within 14 days after receiving the first COVID-19 vaccination, compared with the period before vaccination in England and Wales [81]. Andrejko et al. reported that COVID-19 vaccination within households was associated with increased social contact outside the home for both vaccinated adults and their unvaccinated children [82]. In this study, these trends were confirmed, and the Peltzman effect associated with the interaction between vaccination and the number of newly infected individuals was identified as a new possible cause.

(iv). The correlation coefficients suggest that in segment A, the negative correlation coefficients between the number of newly infected individuals and human flow in each region tended to be higher for the second vaccination than that for the first. However, a test of the difference between the independent correlation coefficients based on the z-distribution showed no significant differences in the correlation coefficients based on the number of vaccinations. Therefore, it is highly likely that the correlation coefficients between the first and second vaccinations were statistically equivalent in each region. For Chiba, the correlation coefficient between the number of newly infected individuals and human flow was statistically uncorrelated when the first vaccination had not started. However, the correlation coefficient between the number of newly infected individuals and human flow was statistically significant when the second round of vaccination

had not started. This suggests that people's awareness of the risk of infection may differ before and after vaccination. In particular, it is possible that the perception of infection risk in terms of the number of newly infected individuals in the respective area was improved by vaccination. This phenomenon is likely attributable to the early dissemination of information by the Ministry of Health, Labour and Welfare of Japan and the mass media regarding the necessity of the second vaccination to achieve a significant preventive effect against infection [83]. It was also communicated that antibody formation takes approximately one to two weeks, during which the incidence rate mirrors that in unvaccinated individuals. It is thought that these statements increased awareness of the risk of infection. Therefore, it is believed that fewer people were bold enough to take action after the first vaccination, and attention to the number of newly infected individuals in their living locations increased.

(v). The correlation coefficients between regions suggested that, in situations where vaccination had not yet begun, people tended to refer to the number of newly infected individuals in the capital, Tokyo, or Japan as a whole, rather than in their home region. However, a test of the difference between the independent correlation coefficients based on the z-distribution showed no significant differences in the correlation coefficients between regions. Therefore, it is highly likely that the correlation coefficients between regions, such as those between Japan and Tokyo and between Tokyo and Chiba, were statistically equivalent, regardless of the number of vaccinations. However, in the case of the first vaccination dose, the relationship between the number of newly infected individuals in Japan and Tokyo and human flow was not statistically shown to be no correlation (see left and center in Figure 6(a)). By contrast, the relationship between the number of newly infected individuals and human flow in Chiba was statistically no correlation (see right in Figure 6(a)). Thus, it is suggested that in situations where vaccination had not commenced, human flow may have been determined based on the number of newly infected individuals not living in their residential area, but rather in the capital city of Tokyo or across Japan as a whole. It is generally believed that people who go out in a situation where infectious diseases are spreading make decisions based on the number of newly infected individuals in an area close to their residence. However, in this study, Narashino citizens referred to the number of newly infected individuals in Tokyo and Japan rather than in Chiba, which is contrary to the general assumption. This is likely related to the locations and broadcast areas of key stations in Japan. The five key stations in Japan are Nippon Television Network Corporation [84], TV Asahi Corporation [85], TBS HOLDINGS, INC. [86], TV TOKYO Corporation [87], and Fuji Television Network, Inc. [88]. Their headquarters are located in Tokyo, the capital of Japan, and their broadcast areas cover the entire country. Thus, information related to the number of newly infected individuals sourced from the mass media often covers all of Japan or Tokyo, the capital of Japan, and the headquarters of key stations. Therefore, it is thought that not only Narashino residents, but also people all over Japan, were in a situation where they often saw the number of newly infected individuals in Japan and Tokyo. Thus, the human flow of Narashino citizens was more strongly influenced by the number of newly infected individuals in Japan and Tokyo than by the number of newly infected individuals in Chiba. If this is correct, these behaviors are likely to cause an increase in the number of newly infected individuals and economic losses before the start of vaccination. For example, if the number of newly infected individuals is high in Chiba but low across Japan and Tokyo, it is likely to lead to an increase in the number of

Table 6. SWOT matrix for this study.

Strengths
The analysis has been conducted with maximum consideration of the characteristics of the dataset, and the reliability of the results has been sufficiently enhanced. Facts reported in previous studies are emphasized, and the importance of factors is presented. In addition, we have presented the differences in human flow with respect to the vaccination status, depending on the number of vaccinations and regional characteristics of the number of newly infected individuals.
Weaknesses
No statistically significant differences in the correlation coefficients were obtained for differences in trends in the human flow by region and number of vaccinations.
Opportunities
The results of this study are consistent with those reported in previous studies, indicating a high degree of validity. It is expected that the hypotheses obtained in this study will be validated in the future.
Threats
Future social advances may make the results and interpretations of this study inapplicable.

newly infected individuals at their place of residence owing to increased daily travel, despite the relatively higher risk of infection. Conversely, if the number of newly infected individuals is high in Japan and Tokyo but low across Chiba, daily travels are suppressed despite the relatively low risk of infection in the living centers, and economic losses are likely to occur in the living centers. Therefore, in situations in which vaccination has not yet begun, it may be important to provide information on the number of newly infected individuals in each region, in addition to information on the number of newly infected individuals nationally and in the capital, to control infection and for economic benefit.

4.2. SWOT analysis of this study

Table 6 shows the pros and cons of this study as a SWOT matrix.

- ***Strengths*** — Among the factors in previous studies identified as being closely related to human flow in the COVID-19 pandemic, this study identifies the factors of particular importance. This was realized by comparing machine-learning models with different algorithms and factor analyzing the model that achieved the best prediction performance. In addition, standardization, grouping, cross-validation, and hyperparameter optimization were applied to these predictive models. In other words, we believe that the analysis has taken into account the dataset's characteristics to the highest extent possible, and that the confidence in the results has been sufficiently increased. The results obtained from this study are a major contribution to the existing knowledge in that they verify the findings reported in previous studies. In addition, the newly presented differences in the importance of the factors and trends in human flow by region and number of vaccinations are considered to be useful for the future development of infectious disease control measures.
- ***Weaknesses*** — In this study, no statistically significant differences in the correlation coefficients were obtained for differences in trends in the vaccination statuses among regions and among vaccination frequencies. Therefore, we believe that strong assertions should be avoided from this study because it is not possible at this time to make any definitive statements about these results. In other

words, it should be noted that these results and interpretations only provide new hypotheses.

- ***Opportunities*** — Compared to the early days of the COVID-19 outbreak, data availability has improved and statistical methods, including machine learning, have been developed. Accordingly, several findings on COVID-19 have been reported, strengthening the scientific evidence for the effectiveness of policies and public health measures. The results of this study are consistent with many of these reports, and we believe that they have demonstrated a high degree of validity. We believe that these developments will continue to realize significant progress in the future. Therefore, new technologies and datasets may allow for more in-depth discussions. We believe that the hypotheses obtained in this study can be tested and solved in the future.

- ***Threats*** — The results and interpretation depend on the input variables and machine-learning model used. Therefore, depending on the variables and algorithms considered, there may be discrepancies in the results and interpretations. In addition, changes in the social, economic, and technological environment may cause changes in the assumptions in previous studies. Specifically, changes in social values and behaviors, policies, and technological innovations are noted. Future social progress including these factors may render the results and interpretations of this study inapplicable.

5. Conclusions, limitations and prospects

In this study, machine learning was used to analyze political and legal factors as well as health and safety factors that are expected to have a strong impact on human flow in the COVID-19 pandemic. Next, the cumulative number of vaccinated individuals and number of newly infected individuals were subdivided by regional characteristics and the number of vaccinations to investigate the detailed impact on human flow through visual and correlational analyses. As a result, the following four possibilities were suggested:

(i). There are complex interactions among the input variables (number of newly infected individuals, number of deaths, number of vaccinated individuals, government infectious disease control measures, rate of work-from-home utilization, month-end hospital bed occupancy rate, day of the week, and climate conditions) employed in this study, and nonlinear relationships exist between these variables and human flows. Tree-based models can take these into account, and LightGBM can effectively reduce bias, thus achieving good forecasting performance.

(ii). Previous studies have reported that the effects of the state-of-emergency and priority measures to prevent the spread of the disease weaken with repeated interventions, and that the implementation of regulations increases risk-compensating behavior [73–75]. In this study, the contribution of government measures against infectious diseases was found to be almost zero (see Table 4), which supports previous findings and newly presents the possibility that their impact is significant. In addition, as a reason why the cumulative number of vaccinated individuals and number of newly infected individuals are highly related to human flow, it was suggested that there is a close relationship between human flow and the Peltzman effect.

(iii). The Peltzman effect may reduce the perception of infection risk even when people experience an increase in infections after vaccination.

(iv). Vaccination may improve people's perception of the risk of infection in their residential areas. This may be due to the influence of the government's and mass media's calls for vaccination and on the perception of infection risk.

(v). Before the start of vaccination, the daily travel distance may be determined by the number of newly infected individuals in the nation or in the capital. This may be related to the fact that the key station in Japan is located in Tokyo, the capital of Japan, and the broadcasting area is nationwide. These behaviors may cause an increase in the number of newly infected individuals and economic losses in the residential areas before the start of vaccination.

In this study, factors with a particularly high association with human flow were identified in the COVID-19 pandemic. In addition, the impact of the regional characteristics of these factors and different effects of the number of vaccinations on human flow were elucidated. Our findings support those of previous studies and provide new insights. The government infectious disease control measures against COVID-19 may have a smaller impact than that of the other factors. In addition, the Peltzman effect, which is caused by the interaction between the number of newly infected individuals and cumulative number of vaccinated individuals, may have a significant impact on the change in human flow. Therefore, it is important to review government policies to effectively control the Peltzman effect in future infectious disease control measures. However, it is important to establish a system to voluntarily suppress mobility, and to avoid suppressing the Peltzman effect by strictly enforced policies. Watanabe and Yabu suggested that the provision of appropriate information to encourage people to change their behavior is more important in controlling the spread of the novel coronavirus than strong and legally binding measures [51]. Moreover, Coccia reported that high levels of strict restriction policies may not be effective measures in controlling the spread and adverse effects of the pandemic. It has been suggested that the average number of confirmed cases and fatality rates associated with COVID-19 tend to be lower in less-restricted countries [33, 89]. Possible mechanisms to voluntarily control mobility include vaccination at appropriate times and cooperation of the mass media. Based on the findings of this study, awareness of COVID-19 can be maintained at a high level by reviewing the timing of vaccination. Lee et al. reported an increase in short-distance travel by private vehicles during the pandemic [90]. However, the mass media can provide the number of newly infected persons throughout Japan and outside the capital city, so that appropriate information on the place of residence can be ascertained. These are parts of the uncoerced suppression of the Peltzman effect, which allows the public to accurately judge the risk of infection. As a result, it is expected that the control of human flow and economic loss will be realized. However, the existence of a good government is a prerequisite for such a realization [91–93]. In many countries, pandemic preparedness is not sufficient, and there is still room for improvement [94]. Therefore, it is important to design long-term prevention strategies for the next pandemic in anticipation of a future epidemic of an infectious disease similar to COVID-19 [95, 96]. During the COVID-19 pandemic and in the presence of new pandemics as well, designing effective policy responses to mitigate the impact of the initial spread phase is one of the fundamental problems [97]. The findings of this study may contribute to solving this problem. However, this study has four limitations:

- The groups in Table 2 were assumed to be independent of one another. Normally, when there is a possible correlation among features, it is necessary to exclude or compress the features by feature selection or other means in advance. This study aimed to clarify not only the disclosed information

on COVID-19 and the effects of implemented measures, but also the effects of the regional characteristics and number of vaccinations. Therefore, it was necessary to use features with possible correlations. By calculating the feature importance by group, we attempted to stabilize the learning of prediction models and reduce the bias in the group feature importance. However, the analysis was performed under the condition that the bias was not completely eliminated and that it had some negative effects.

- The cause-and-effect assumption was that human flow varies according to the groups in Table 2. In the real world, the features and response variable in this study are thought to affect one another in complexity. However, it is challenging to conduct an analysis after unraveling all of these factors. Therefore, the results obtained in this study are trends that are confirmed when limited to these assumptions and do not necessarily correspond to trends in the real world.
- The analysis was limited to the factors in Table 2, despite the existence of countless factors that affect human flow in the COVID-19 pandemic. To conduct a more comprehensive factor analysis in predicting human flow, it is considered necessary to add factors that may have even a slight influence on human flow to the verification target. However, it is difficult to conduct an analysis that considers all of these factors. This is because there are factors that make it difficult to satisfy the quality and quantity of data available for the study adequately owing to the fact that real data were used for the analysis. In addition, the sample size must be much larger than the features for proper use of machine-learning models. Thus, this study employed 32 factors for which the data were reliable and could be analyzed with the available sample size. However, to achieve higher prediction performance and conduct a factor analysis that is closer to the real world, it is necessary to increase both the number of features and sample size [98]. This is recognized as a future challenge.
- The results obtained from the analysis may be biased toward a specific population because the analysis was limited to Narashino City, Japan. Further verification for many countries and locations is necessary for a generalized discussion. In the future, it will be necessary to determine whether these results are specific to Japan or if similar trends are observed in other countries and locations, and to identify any commonalities.

To complement these limitations, future studies should first focus on enhancing data collection, including samples and features. Second, the results and hypotheses obtained in this study need to be examined in further detail by employing different modeling approaches. Specific methods that can be considered include regime switching model approaches [99], multivariate adaptive regression splines [100], fuzzy-regression approaches [101, 102], and closed-loop supply chain networks [103] that take into account uncertainty. These methods are applied in various analyses, including the characteristics of pandemic peaks and troughs [104], analyzing factors that influence vaccine perception [105], evaluating the causal impact of COVID-19 vaccination on the reduction of preventive health behaviors [106], and designing optimal delivery and retrieval of medical equipment while accounting for the uncertainties caused by COVID-19 [107]. The application of these mathematical methods is expected to provide new insights in health care during various infectious diseases.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported in part by the JSPS Grant-in-Aid for Scientific Research (C) (Grant No. 21K04535) and JSPS Grant-in-Aid for Early-Career Scientists (Grant No. 23K13517).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. V. Saladino, D. Algeri, V. Auriemma, The psychological and social impact of Covid-19: new perspectives of well-being, *Front. Psychol.*, (2020), 2550. https://doi.org/10.3389/fpsyg.2020.577684
2. S. Shanbehzadeh, M. Tavahomi, N. Zanjari, I. Ebrahimi-Takamjani, S. Amiri-Arimi, Physical and mental health complications post-COVID-19: Scoping review, *J. Psychosom. Res.*, **147** (2021), 110525. https://doi.org/10.1016/j.jpsychores.2021.110525
3. I. Ali, O. M. Alharbi, COVID-19: Disease, management, treatment, and social impact, *Sci. Total Environ.*, **728** (2020), 138861. https://doi.org/10.1016/j.scitotenv.2020.138861
4. S. Pokhrel, R. Chhetri, A literature review on impact of COVID-19 pandemic on teaching and learning, *Higher Educ. Future*, **8** (2021), 133–141. https://doi.org/10.1177/2347631120983481
5. A. D. Kaye, C. N. Okeagu, A. D. Pham, R. A. Silva, J. J. Hurley, B. L. Arron, et al., Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives, *Best Pract. Res. Clin. Anaesthesiol.*, **35** (2021), 293–306. https://doi.org/10.1016/j.bpa.2020.11.009
6. S. Naseer, S. Khalid, S. Parveen, K. Abbass, H. Song, M. V. Achim, COVID-19 outbreak: Impact on global economy, *Front. Public Health*, **10** (2023), 1009393. https://doi.org/10.3389/fpubh.2022.1009393
7. M. Coccia, Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID, *Sci. Total Environ.*, **729** (2020), 138474. https://doi.org/10.1016/j.scitotenv.2020.138474
8. M. Coccia, How do low wind speeds and high levels of air pollution support the spread of COVID-19?, *Atmos. Pollut. Res.*, **12** (2021), 437–445. https://doi.org/10.1016/j.apr.2020.10.002
9. M. Coccia, The effects of atmospheric stability with low wind speed and of air pollution on the accelerated transmission dynamics of COVID-19, *Int. J. Environ. Stud.*, **78** (2021), 1–27. https://doi.org/10.1080/00207233.2020.1802937
10. E. Bontempi, M. Coccia, International trade as critical parameter of COVID-19 spread that outclasses demographic, economic, environmental, and pollution factors, *Environ. Res.*, **201** (2021), 111514. https://doi.org/10.1016/j.envres.2021.111514

11. E. Bontempi, M. Coccia, S. Vergalli, A. Zanoletti, Can commercial trade represent the main indicator of the COVID-19 diffusion due to human-to-human interactions? A comparative analysis between Italy, France, and Spain, *Environ. Res.*, **201** (2021), 111529. https://doi.org/10.1016/j.envres.2021.111529

12. Y. Diao, S. Kodera, D. Anzai, J. Gomez-Tames, E. A. Rashed, A. Hirata, Influence of population density, temperature, and absolute humidity on spread and decay durations of COVID-19: A comparative study of scenarios in China, England, Germany, and Japan, *One Health*, **12** (2021), 100203. https://doi.org/10.1016/j.onehlt.2020.100203

13. C. Magazzino, M. Mele, M. Coccia, A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality, *Epidemiol. Infect.*, **150** (2022), e168. https://doi.org/10.1017/S0950268822001418

14. A. Nunez-Delgado, E. Bontempi, M. Coccia, M. Kumar, J. L. Domingo, SARS-CoV-2 and other pathogenic microorganisms in the environment, *Environ. Res.*, **201** (2021), 111606. https://doi.org/10.1016/j.envres.2021.111606

15. M. Coccia, COVID-19 pandemic over 2020 (with lockdowns) and 2021 (with vaccinations): similar effects for seasonality and environmental factors, *Environ. Res.*, **208** (2022), 112711. https://doi.org/10.1016/j.envres.2022.112711

16. H. Kato, A. Takizawa, Human mobility and infection from Covid-19 in the Osaka metropolitan area, *npj Urban Sustainability*, **2** (2022), 20. https://doi.org/10.1038/s42949-022-00066-w

17. Y. Nohara, T. Manabe, Impact of human mobility and networking on spread of COVID-19 at the time of the 1st and 2nd epidemic waves in Japan: An effective distance approach, *PLoS One*, **17** (2022), e0272996. https://doi.org/10.1371/journal.pone.0272996

18. A. Lison, J. Persson, N. Banholzer, S. Feuerriegel, Estimating the effect of mobility on SARS-CoV-2 transmission during the first and second wave of the COVID-19 epidemic, Switzerland, March to December 2020, *Eurosurveillance*, **27** (2022), 2100374. https://doi.org/10.2807/1560-7917.ES.2022.27.10.2100374

19. K. Hibiya, A. Shinzato, H. Iwata, T. Kinjo, M. Tateyama, K. Yamamoto, et al., Effect of voluntary human mobility restrictions on vector-borne diseases during the COVID-19 pandemic in Japan: A descriptive epidemiological study using a national database (2016 to 2021), *PLoS One*, **18** (2023), e0285107. https://doi.org/10.1371/journal.pone.0285107

20. M. Zhang, S. Wang, T. Hu, X. Fu, X. Wang, Y. Hu, et al., Human mobility and COVID-19 transmission: a systematic review and future directions, *Ann. GIS*, **28** (2022), 501–514. https://doi.org/10.1080/19475683.2022.2041725

21. N. Askitas, K. Tatsiramos, B. Verheyden, Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study, *Sci. Rep.*, **11** (2021), 1972. https://doi.org/10.1038/s41598-021-81442-x

22. R. S. John, J. C. Miller, R. L. Muylaert, D. T. Hayman, High connectivity and human movement limits the impact of travel time on infectious disease transmission, *J. R. Soc. Interface*, **21** (2024), 20230425. https://doi.org/10.1098/rsif.2023.0425

23. J. Li, C. Zhuang, W. Zou, A tale of lockdown policies on the transmission of COVID-19 within and between Chinese cities: A study based on heterogeneous treatment effect, *Econ. Hum. Biol.*, **53** (2024), 101365. https://doi.org/10.1016/j.ehb.2024.101365

24. M. G. Thompson, J. L. Burgess, A. L. Naleway, H. Tyner, S. K. Yoon, J. Meece, et al., Prevention and attenuation of Covid-19 with the BNT162b2 and mRNA-1273 vaccines, *N. Engl. J. Med.*, **385** (2021), 320–329. https://doi.org/10.1056/NEJMoa2107058

25. V. Hall, S. Foulkes, F. Insalata, P. Kirwan, A. Saei, A. Atti, et al., Protection against SARS-CoV-2 after Covid-19 vaccination and previous infection, *N. Engl. J. Med.*, **386** (2022), 1207–1220. https://doi.org/10.1056/NEJMoa2118691

26. S. J. Thomas, E. D. Moreira, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, et al., Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months, *N. Engl. J. Med.*, **385** (2021), 1761–1773. https://doi.org/10.1056/NEJMoa2110345

27. J. Sadoff, G. Gray, A. Vandebosch, V. Cárdenas, G. Shukarev, B. Grinsztejn, et al., Safety and efficacy of single-dose Ad26. COV2. S vaccine against Covid-19, *N. Engl. J. Med.*, **384** (2021), 2187–2201. https://doi.org/10.1056/NEJMoa2101544

28. Y. Z. Huang, C. C. Kuan, Vaccination to reduce severe COVID-19 and mortality in COVID-19 patients: a systematic review and meta-analysis, *Eur. Rev. Med. Pharmacol. Sci.*, **26** (2022). https://doi.org/10.26355/eurrev_202203_28248

29. B. Trogen, A. Caplan, Risk compensation and COVID-19 vaccines, *Ann. Intern. Med.*, **174** (2021), 858–859. https://doi.org/10.7326/M20-8251

30. K. P. Iyengar, P. Ish, R. Botchu, V. K. Jain, R. Vaishya, Influence of the Peltzman effect on the recurrent COVID-19 waves in Europe, *Postgrad. Med. J.*, **98** (2022), e110–e111. https://doi.org/10.1136/postgradmedj-2021-140234

31. J. Guo, C. Deng, F. Gu, Vaccinations, mobility and COVID-19 transmission, *Int. J. Environ. Res. Public Health*, **19** (2021), 97. https://doi.org/10.3390/ijerph19010097

32. L. L. Liang, H. M. Le, C. Y. Wu, C. Y. Sher, A. McGuire, Human mobility increased with vaccine coverage and attenuated the protection of COVID-19 vaccination: A longitudinal study of 107 countries, *J. Glob. Health*, **13** (2023). https://doi.org/10.7189/jogh.13.06009

33. M. Coccia, Improving preparedness for next pandemics: Max level of COVID-19 vaccinations without social impositions to design effective health policy and avoid flawed democracies, *Environ. Res.*, **213** (2022), 113566. https://doi.org/10.1016/j.envres.2022.113566

34. H. Barbosa, S. Hazarie, B. Dickinson, A. Bassolas, A. Frank, H. Kautz, et al., Uncovering the socioeconomic facets of human mobility, *Sci. Rep.*, **11** (2021), 8616. https://doi.org/10.1038/s41598-021-87407-4

35. C. Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, et al., Analyzing and geo-visualizing individual human mobility patterns using mobile call records, in *Proceedings of the 18th International Conference on Geoinformatics*, (2010), 1–7. https://doi.org/10.1109/GEOINFORMATICS.2010.5567857

36. A. Paez, F. A. Lopez, T. Menezes, R. Cavalcanti, M. G. d. R. Pitta, A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain, *Geogr. Anal.*, **53** (2021), 397–421. https://doi.org/10.1111/gean.12241

37. T. Chowdhury, H. Chowdhury, E. Bontempi, M. Coccia, H. Masrur, S. M. Sait, et al., Are mega-events super spreaders of infectious diseases similar to COVID-19? A look into Tokyo 2020 Olympics and Paralympics to improve preparedness of next international events, *Environ. Sci. Pollut. Res.*, **30** (2023), 10099–10109. https://doi.org/10.1007/s11356-022-22660-2

38. M. Murakami, K. Fujii, W. Naito, M. Kamo, M. Kitajima, T. Yasutaka, et al., COVID-19 infection risk assessment and management at the Tokyo 2020 Olympic and Paralympic Games: A scoping review, *J. Infect. Public Health*, **17** (2024), 18–26. https://doi.org/10.1016/j.jiph.2023.03.025

39. S. Hu, C. Xiong, M. Yang, H. Younes, W. Luo, L. Zhang, A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic, *Transp. Res. Part C: Emerg. Technol.*, **124** (2021), 102955. https://doi.org/10.1016/j.trc.2020.102955

40. D. Nakamoto, S. Nojiri, C. Taguchi, Y. Kawakami, S. Miyazawa, M. Kuroki, et al., The impact of declaring the state of emergency on human mobility during COVID-19 pandemic in Japan, *Clin. Epidemiol. Glob. Health*, **17** (2022), 101149. https://doi.org/10.1016/j.cegh.2022.101149

41. M. Chakraborty, M. S. Mahmud, T. J. Gates, S. Sinha, Analysis and prediction of human mobility in the United States during the early stages of the COVID-19 pandemic using regularized linear models, *Transp. Res. Rec.*, **2677** (2023), 380–395. https://doi.org/10.1177/03611981211067794

42. M. U. G. Kraemer, C. H. Yang, B. Gutierrez, C. H. Wu, B. Klein, D. M. Pigott, et al., The effect of human mobility and control measures on the COVID-19 epidemic in China, *Science*, **368** (2020), 493–497. https://doi.org/10.1126/science.abb4218

43. H. Fang, L. Wang, Y. Yang, Human mobility restrictions and the spread of the novel coronavirus (2019-nCoV) in China, *J. Public Econ.*, **191** (2020), 104272. https://doi.org/10.1016/j.jpubeco.2020.104272

44. J. Y. L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z. W. Hong, et al., Mitigating the multicollinearity problem and its machine learning approach: a review, *Mathematics*, **10** (2022), 1283. https://doi.org/10.3390/math10081283

45. C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, *BMC Bioinf.*, **9** (2008), 1–11. https://doi.org/10.1186/1471-2105-9-307

46. *Prime Minister's Office of Japan*, Prime minister's office of Japan. Available from: https://japan.kantei.go.jp.

47. *Ministry of Health, Labour and Welfare of Japan*, Welcome to ministry of health, labour and welfare. Available from: https://www.mhlw.go.jp/english.

48. *Tokyo Metropolitan Government*, Home-tokyo metropolitan government. Available from: https://www.metro.tokyo.lg.jp/english/index.html.

49. *Japan Meteorological Agency*, Japan meteorological agency. Available from: https://www.jma.go.jp/jma/indexe.html.

50. *Portal Site of Official Statistics of Japan*, Portal site of official statistics of Japan. Available from: https://www.e-stat.go.jp/en.

51. T. Watanabe, T. Yabu, Japan's voluntary lockdown, *PLoS One*, **16** (2021), e0252468. https://doi.org/10.1371/journal.pone.0252468

52. Q. He, Z. Zhang, Y. Xie, The impact of COVID-19 on Americans' attitudes toward China: Does local incidence rate matter?, *Soc. Psychol. Q.*, **85** (2022), 84–107. https://doi.org/10.1177/01902725211072773

53. X. Chen, X. Di, How the covid-19 pandemic influences human mobility? similarity analysis leveraging social media data, in *Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, (2022), 2955–2960. https://doi.org/10.1109/ITSC55140.2022.9922060

54. J. Yuan, M. Li, G. Lv, Z. K. Lu, Monitoring transmissibility and mortality of COVID-19 in Europe, *Int. J. Infect. Dis.*, **95** (2020), 311–315. https://doi.org/10.1016/j.ijid.2020.03.050

55. H. Ito, N. Kawazoe, Examining transportation mode changes during COVID-19 in Toyama, Japan, *Reg. Stud. Reg. Sci.*, **10** (2023), 253–272. https://doi.org/10.1080/21681376.2023.2180425

56. Y. Hara, H. Yamaguchi, Japanese travel behavior trends and change under COVID-19 state-of-emergency declaration: Nationwide observation by mobile phone location data, *Transp. Res. Interdiscip. Perspect.*, **9** (2021), 100288. https://doi.org/10.1016/j.trip.2020.100288

57. F. Benita, Human mobility behavior in COVID-19: A systematic literature review and bibliometric analysis, *Sustainable Cities Soc.*, **70** (2021), 102916. https://doi.org/10.1016/j.scs.2021.102916

58. E. Thuillier, L. Moalic, S. Lamrous, A. Caminada, Clustering weekly patterns of human mobility through mobile phone data, *IEEE Trans. Mobile Comput.*, **17** (2017), 817–830. https://doi.org/10.1109/TMC.2017.2742953

59. T. Mahajan, G. Singh, G. Bruns, An experimental assessment of treatments for cyclical data, in *Proceedings of the 2021 Computer Science Conference for CSU Undergraduates, Virtual*, **6** (2021), 22.

60. T. Horanont, S. Phithakkitnukoon, T. W. Leong, Y. Sekimoto, R. Shibasaki, Weather effects on the patterns of people's everyday activities: a study using GPS traces of mobile phone users, *PLoS One*, **8** (2013), e81153. https://doi.org/10.1371/journal.pone.0081153

61. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **68** (2006), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

62. J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and a sparse group lasso, preprint, arXiv:1001.0736. https://doi.org/10.48550/arXiv.1001.0736

63. N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse-group lasso, *J. Comput. Graph. Stat.*, **22** (2013), 231–245. https://doi.org/10.1080/10618600.2012.681250

64. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324

65. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, **12** (2011), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

66. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.*, **30** (2017).

67. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (2019), 2623–2631. https://doi.org/10.1145/3292500.3330701

68. B. Gregorutti, B. Michel, P. Saint-Pierre, Grouped variable importance with random forests and application to multiple functional data analysis, *Comput. Stat. Data Anal.*, **90** (2015), 15–35. https://doi.org/10.1016/j.csda.2015.04.002

69. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, **2** (2009). https://doi.org/10.1007/978-0-387-21606-5

70. S. Luo, T. Chen, Two derivative algorithms of gradient boosting decision tree for silicon content in blast furnace system prediction, *IEEE Access*, **8** (2020), 196112–196122. https://doi.org/10.1109/ACCESS.2020.3034566

71. X. Zeng, Length of stay prediction model of indoor patients based on light gradient boosting machine, *Comput. Intell. Neurosci.*, 2022. https://doi.org/10.1155/2022/9517029

72. Y. Hu, Z. Sun, Y. Han, W. Li, L. Pei, Evaluate pavement skid resistance performance based on Bayesian-LightGBM using 3D surface macrotexture data, *Materials*, **15** (2022), 5275. https://doi.org/10.3390/ma15155275

73. S. Okamoto, State of emergency and human mobility during the COVID-19 pandemic in Japan, *J. Transp. Health*, **26** (2022), 101405. https://doi.org/10.1016/j.jth.2022.101405

74. S. Fukui, Long-term changes in human mobility responses to COVID-19-related information in Japan, *medRxiv*, (2022), 2022–08. https://doi.org/10.1101/2022.08.15.22278703

75. K. Henk, F. Rosing, F. Wolff, S. B. Frenzel, R. van Dick, V. A. Erkens, et al., An examination and extension of the Peltzman effect during the Covid-19 pandemic, *Curr. Res. Ecol. Soc. Psychol.*, **4** (2023), 100091. https://doi.org/10.1016/j.cresp.2023.100091

76. S. Peltzman, The effects of automobile safety regulation, *J. Polit. Econ.*, **83** (1975), 677–725. https://doi.org/10.1086/260352

77. R. O. Nanda, A. A. Nursetyo, A. L. Ramadona, M. A. Imron, A. Fuad, A. Setyawan, et al., Community mobility and COVID-19 dynamics in Jakarta, Indonesia, *Int. J. Environ. Res. Public Health*, **19** (2022), 6671. https://doi.org/10.3390/ijerph19116671

78. M. Coccia, COVID-19 vaccination is not a sufficient public policy to face crisis management of next pandemic threats, *Public Organ. Rev.*, **23** (2023), 1353–1367. https://doi.org/10.1007/s11115-022-00661-6

79. M. Shibamoto, S. Hayaki, Y. Ogisu, COVID-19 infection spread and human mobility, *J. Jap. Int. Econ.*, **64** (2022), 101195. https://doi.org/10.1016/j.jjie.2022.101195

80. H. Kwon, C. Koylu, Revealing associations between spatial time series trends of COVID-19 incidence and human mobility: an analysis of bidirectionality and spatiotemporal heterogeneity, *Int. J. Health Geogr.*, **22** (2023), 33. https://doi.org/10.1186/s12942-023-00357-0

81. A. Serisier, S. Beale, Y. Boukari, S. Hoskins, V. Nguyen, T. Byrne, et al., A case-crossover study of the effect of vaccination on SARS-CoV-2 transmission relevant behaviours during a period of national lockdown in England and Wales, *Vaccine*, **41** (2023), 511–518. https://doi.org/10.1016/j.vaccine.2022.11.073

82. K. L. Andrejko, J. R. Head, J. A. Lewnard, J. V. Remais, Longitudinal social contacts among school-aged children during the COVID-19 pandemic: the Bay Area Contacts among Kids (BACK) study, *BMC Infect. Dis.*, **22** (2022), 242. https://doi.org/10.1186/s12879-022-07218-4

83. *Ministry of Health, Labour and Welfare of Japan*, COVID-19 Vaccine Q&A. Available from: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/vaccine_qa_archive.html#effect_4.

84. *Nippon Television Network Corporation*, Corporate data. Available from: https://www.ntv.co.jp/english/an/cd.html.

85. *TV Asahi Corporation*, Corporate profile. Available from: https://company.tv-asahi.co.jp/e/profile/index.html.

86. *TBS HOLDINGS, INC.*, Corporate profile. Available from: https://www.tbsholdings.co.jp/en/about/corporate/companyprofile.html.

87. *TV TOKYO Corporation*, Corporate profile. Available from: https://www.tv-tokyo.co.jp/kaisha/company/profile.html.

88. *Fuji Television Network, Inc.*, Corporate profile. Available from: https://www.fujitv.com/about/corporate_profile.

89. M. Coccia, Effects of strict containment policies on COVID-19 pandemic crisis: lessons to cope with next pandemic impacts, *Environ. Sci. Pollut. Res.*, **30** (2023), 2020–2028. https://doi.org/10.1007/s11356-022-22024-w

90. S. Lee, E. Ko, K. Jang, S. Kim, Understanding individual-level travel behavior changes due to COVID-19: Trip frequency, trip regularity, and trip distance, *Cities*, **135** (2023), 104223. https://doi.org/10.1016/j.cities.2023.104223

91. I. Benati, M. Coccia, Global analysis of timely COVID-19 vaccinations: improving governance to reinforce response policies for pandemic crises, *Int. J. Health Gov.*, **27** (2022), 240–253. https://doi.org/10.1108/IJHG-07-2021-0072

92. M. Coccia, Optimal levels of vaccination to reduce COVID-19 infected individuals and deaths: A global analysis, *Environ. Res.*, **204** (2022), 112314. https://doi.org/10.1016/j.envres.2021.112314

93. M. Coccia, I. Benati, Negative effects of high public debt on health systems facing pandemic crisis: Lessons from COVID-19 in Europe to prepare for future emergencies, *AIMS Public Health*, **11** (2024), 477–498. https://doi.org/10.3934/publichealth.2024024

94. M. Coccia, Preparedness of countries to face COVID-19 pandemic crisis: strategic positioning and factors supporting effective strategies of prevention of pandemic threats, *Environ. Res.*, **203** (2022), 111678. https://doi.org/10.1016/j.envres.2021.111678

95. M. Coccia, Pandemic prevention: lessons from COVID-19, *Encyclopedia*, **1** (2021), 36. https://doi.org/10.3390/encyclopedia1020036

96. M. Coccia, Sources, diffusion and prediction in COVID-19 pandemic: lessons learned to face next health emergency, *AIMS Public Health*, **10** (2023), 145. https://doi.org/10.3934/publichealth.2023012

97. I. Benati, M. Coccia, Effective contact tracing system minimizes COVID-19 related infections and deaths: policy lessons to reduce the impact of future pandemic diseases, *J. Public Admin. Govern.*, **12** (2022). https://doi.org/10.5296/jpag.v12i2.19834

98. S. Khalilpourazari, H. H. Doulabi, A. Ö. Çiftçioğlu, G. W. Weber, Gradient-based grey wolf optimizer with Gaussian walk: Application in modelling and prediction of the COVID-19 pandemic, *Expert Syst. Appl.*, **177** (2021), 114920. https://doi.org/10.1016/j.eswa.2021.114920

99. E. Savku, N. Azevedo, G. Weber, Optimal control of stochastic hybrid models in the framework of regime switches, in *Modeling, Dynamics, Optimization and Bioeconomics II: DGS III, Porto, Portugal, February 2014, and Bioeconomy VII, Berkeley, USA, March 2014-Selected Contributions 3*, Springer, (2014), 371–387. https://doi.org/10.1007/978-3-319-55236-1_18

100. A. Özmen, E. Kropat, G. W. Weber, Robust optimization in spline regression models for multi-model regulatory networks under polyhedral uncertainty, *Optimization*, **66** (2017), 2135–2155. https://doi.org/10.1080/02331934.2016.1209672

101. E. Kropat, A. Özmen, G. W. Weber, S. Meyer-Nieberg, O. Defterli, Fuzzy prediction strategies for gene-environment networks–Fuzzy regression analysis for two-modal regulatory systems, *RAIRO-Oper. Res.*, **50** (2016), 413–435. https://doi.org/10.1051/ro/2015044

102. E. Kropat, G. W. Weber, Fuzzy target-environment networks and fuzzy-regression approaches, *Numer. Algebra Control Optim.*, **8** (2018), 135–155. https://doi.org/10.3934/naco.2018008

103. E. Özceylan, T. Paksoy, Fuzzy multi-objective linear programming approach for optimising a closed-loop supply chain network, *Int. J. Prod. Res.*, **51** (2013), 2443–2461. https://doi.org/10.1080/00207543.2012.740579

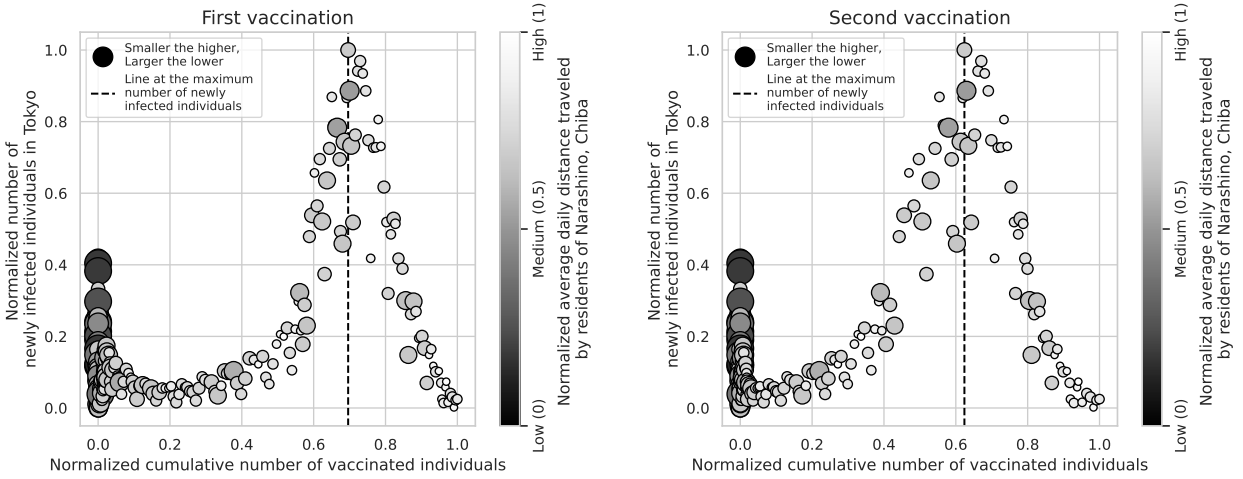
104. P. Haimerl, T. Hartl, Modeling COVID-19 infection rates by regime-switching unobserved components models, *Econometrics*, **11** (2023), 10. https://doi.org/10.3390/econometrics11020010

105. R. Ananda, L. Harsyiah, M. Alfian, Classification of perceptions of the Covid-19 vaccine using multivariate adaptive regression spline, *J. Varian*, **6** (2023), 137–148. https://doi.org/10.30812/varian.v6i2.2639

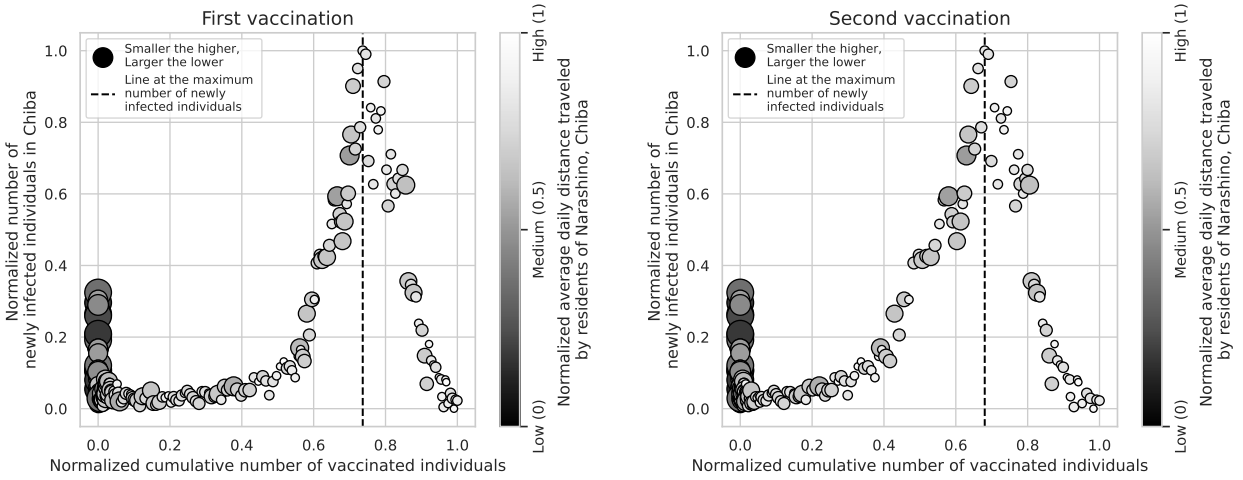
106. F. Chen, H. Nakanishi, Y. Sekizawa, S. Ochi, M. So, Investigating the causal effects of COVID-19 vaccination on the adoption of protective behaviors in Japan: Insights from a fuzzy regression discontinuity design, *PLoS One*, **19** (2024), e0305043. https://doi.org/10.1371/journal.pone.0305043

107. A. Szmelter-Jarosz, J. Ghahremani-Nahr, H. Nozari, A neutrosophic fuzzy optimisation model for optimal sustainable closed-loop supply chain network during COVID-19, *J. Risk Financ. Manage.*, **14** (2021), 519. https://doi.org/10.3390/jrfm14110519

Appendix

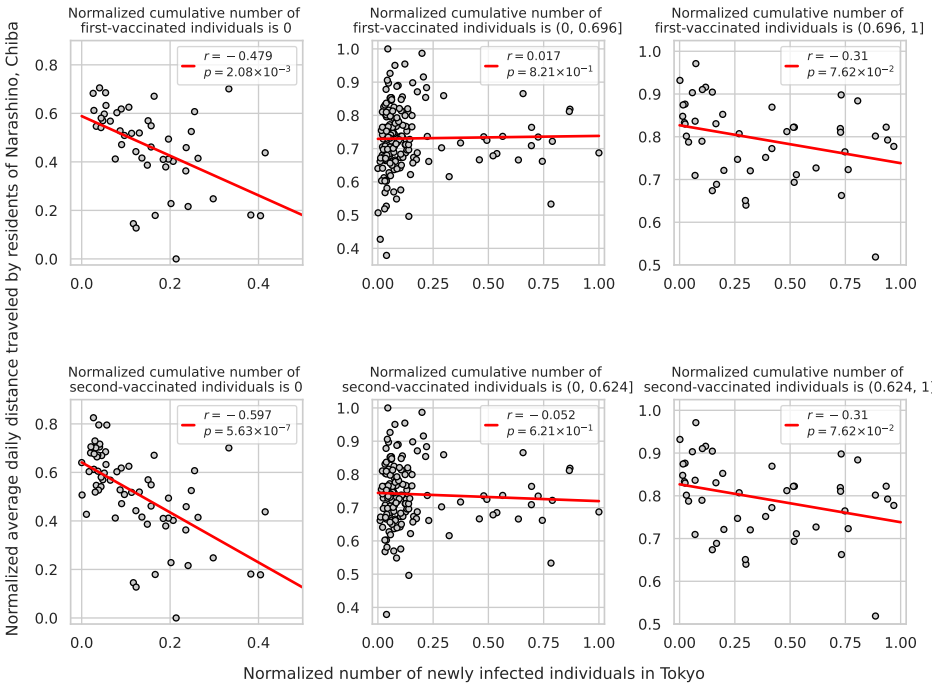


(a) Normalized number of newly infected individuals in Tokyo

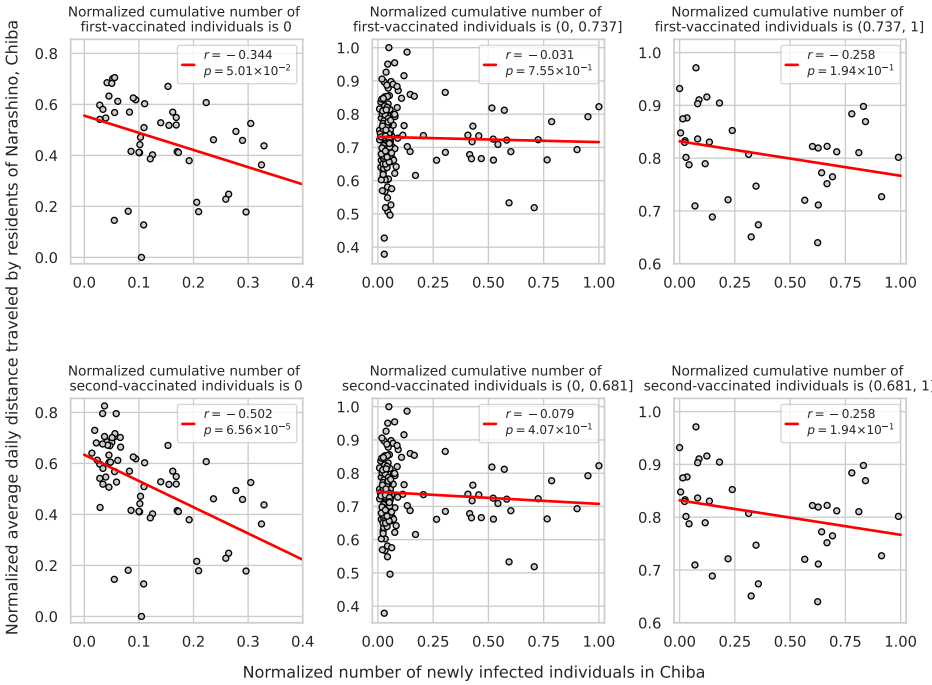


(b) Normalized number of newly infected individuals in Chiba

Figure A1. Scatter plots of the normalized number of newly infected individuals by region, normalized cumulative number of vaccinated individuals by the number of vaccinations, and normalized average daily distance traveled.



(a) Normalized number of newly infected individuals in Tokyo



(b) Normalized number of newly infected individuals in Chiba

Figure A2. Scatter plots, correlations, and adjusted p-values of the normalized number of newly infected individuals, normalized average daily distance traveled by the number of vaccinations, and normalized cumulative number of vaccinated individuals by segment (left: segment A, middle: segment B, right: segment C).

Figure A1(a) shows scatter plots consisting of the number of newly infected individuals in Tokyo, cumulative number of vaccinated individuals subdivided by the number of vaccinations, and human flow of Narashino residents. Figure A1(b) shows scatter plots of similar data for Chiba. In addition, scatter plots of the number of newly infected individuals in Tokyo and human flow divided into segments A, B, and C, along with their correlation coefficients r and adjusted p-values, are shown in Figure A2(a). Scatter plots of the number of newly infected individuals in Chiba and human flow divided into segments A, B, and C, along with their correlation coefficients r and adjusted p-values, are shown in Figure A2(b). For uniformity, all values were normalized using [0, 1] scaling. The three segments were the normalized cumulative number of vaccinated individuals = 0 (segment A), the normalized cumulative number of vaccinated individuals > 0 but below the threshold value (segment B), and the normalized cumulative number of vaccinated individuals > the threshold value but below 1 (segment C). The threshold value refers to the cumulative number of vaccinated individuals when the normalized number of newly infected individuals is at its maximum in the respective scatter plots. The normalized cumulative number of vaccinated individuals at the first vaccination dose was 0.696 and the normalized cumulative number of vaccinated individuals at the second vaccination dose was 0.624 when the number of newly infected individuals in Tokyo was at its maximum (see the dashed line in Figure A1(a)). The normalized cumulative number of vaccinated individuals at the first vaccination dose was 0.737 and the normalized cumulative number of vaccinated individuals at the second vaccination dose was 0.681 when the number of newly infected individuals in Chiba reached its maximum (see the dashed line in Figure A1(b)).

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)