Research Projects

Yufan Zheng

Email: zhjpre@gmail.com / ivan.zane1999@gmail.com

Website: https://yufanzheng.github.io/

Contents

Contents

Epidemic Intervention Evaluation based on Dynamic Systems

Research Background and Significance

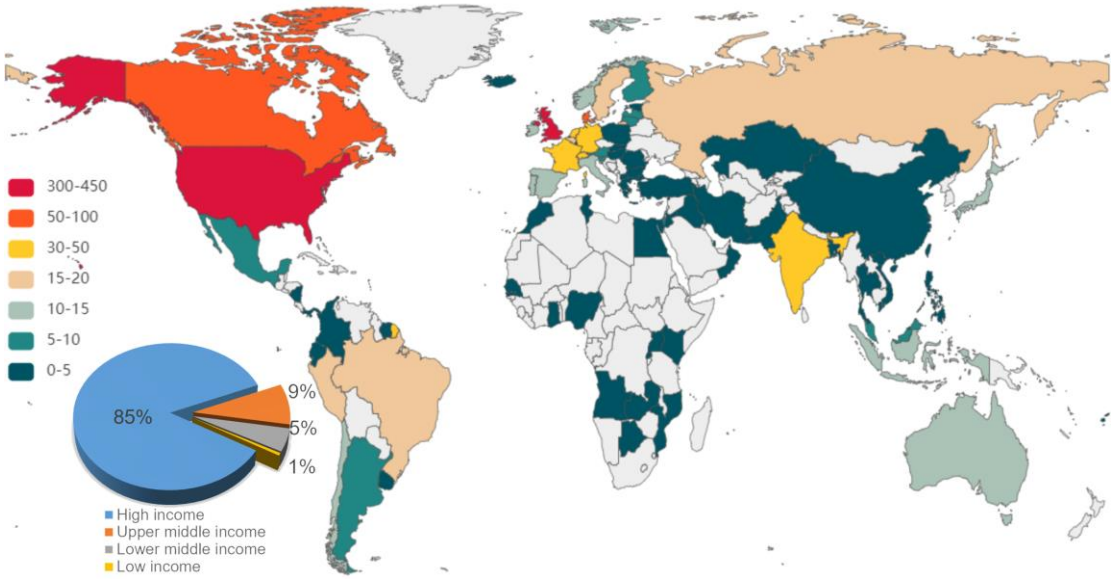


Fig. The number of variants found in the world by April 19, 2022.

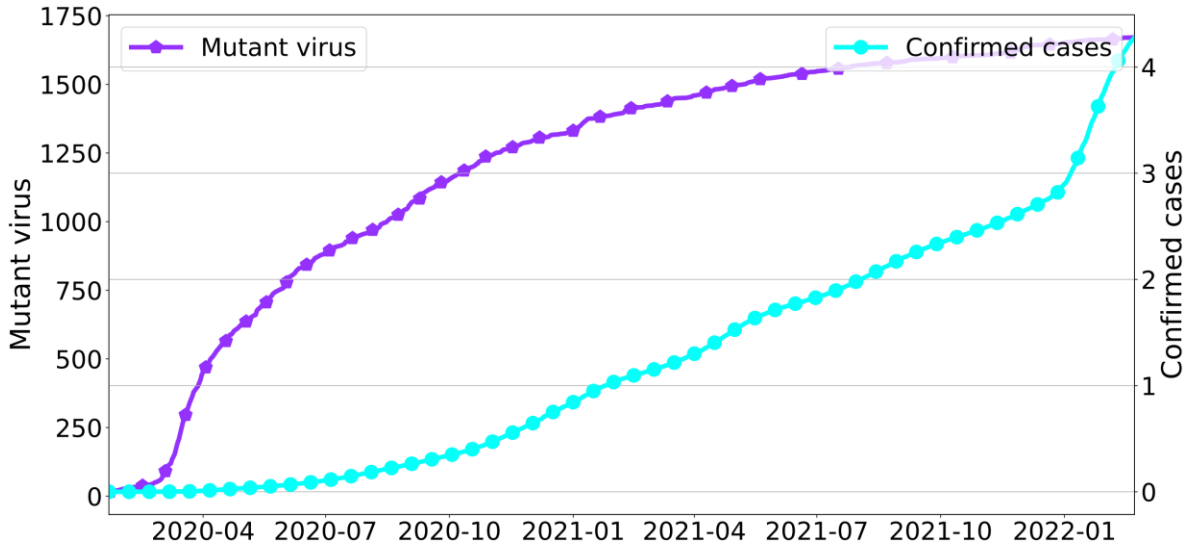


Fig. The cumulative confirmed cases and recorded variants around the world.

Motivation: Multiple-variants coexistence; Viral mutation; Different interventions.

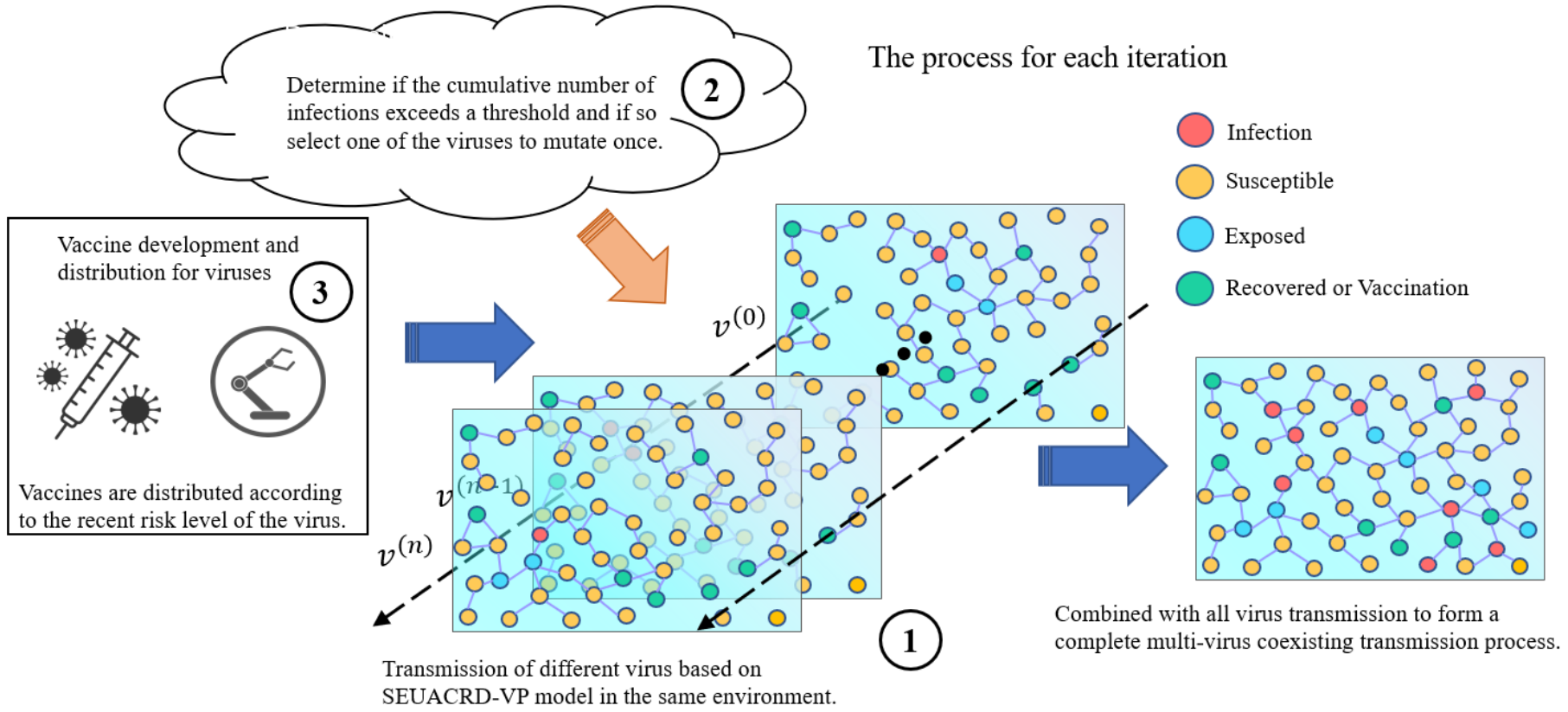
Research aim: To assess the effects of different interventions in the context of viral mutation and multi-variant coexistence.

Achievement:

- Zhan C, **Zheng Y**, Shao L, Chen G, Zhang H. (2023). Modeling the spread dynamics of multiple-variant coronavirus disease under public health interventions: A general framework. Information Sciences, 628, 469-487. (**JCR Q1**)

Methods: framework

1. Proposed a single-virus dynamics model based on the COVID-19 pandemic.
2. Constructed the mechanism for variant mutation and vaccination.
3. Extended the single-virus dynamics model to multi-variant.
4. Estimated the impact of intervention measures.



Methods: Epidemiological model

Proposed a single-virus dynamics model based on the COVID-19 pandemic.

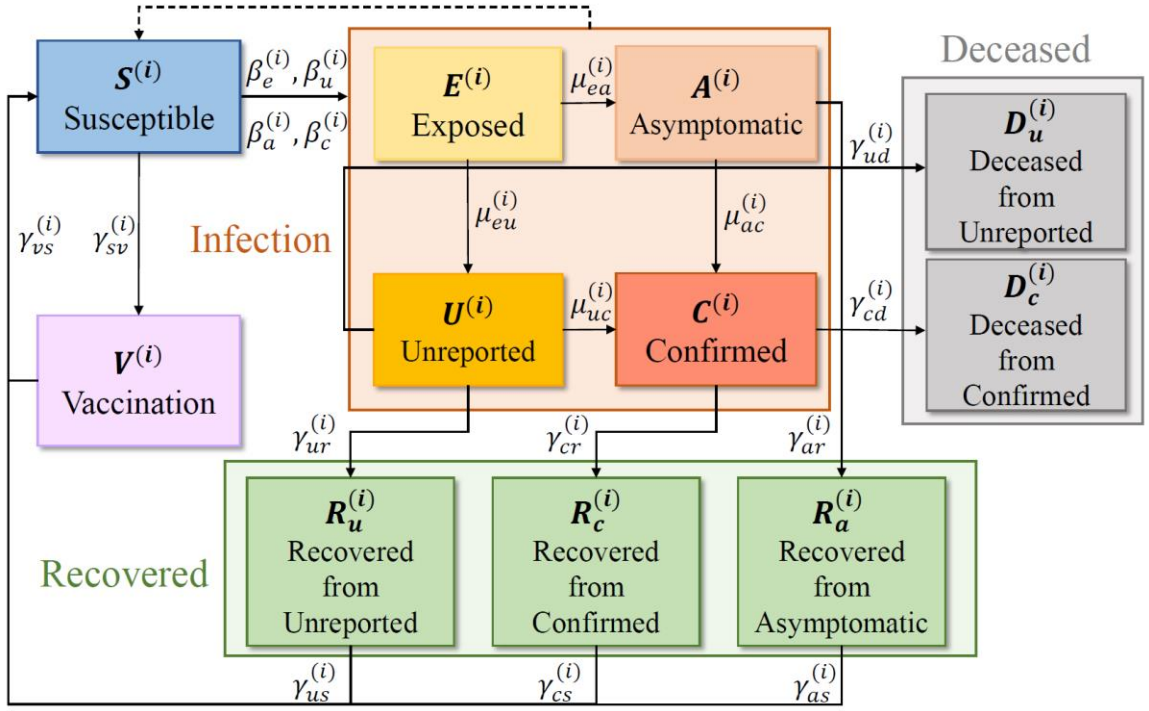


Fig. Flow chart of epidemic model.

$$S^{(i)}(t+1) = S^{(i)}(t) - \frac{(\alpha_e^{(i)}E^{(i)}(t)+\alpha_u^{(i)}U^{(i)}(t)+\alpha_c^{(i)}C^{(i)}(t)+\alpha_a^{(i)}A^{(i)}(t))S^{(i)}(t)}{N_p^{(i)}(t)} + \gamma_{rs}^{(i)}(R_c^{(i)}(t)+R_u^{(i)}(t)+R_a^{(i)}(t)) + \gamma_{vs}^{(i)}V^{(i)}(t) - n_V^{(i)}(t),$$

$$E^{(i)}(t+1) = E^{(i)}(t) + \frac{(\alpha_e^{(i)}E^{(i)}(t)+\alpha_u^{(i)}U^{(i)}(t)+\alpha_c^{(i)}C^{(i)}(t)+\alpha_a A^{(i)}(t))S^{(i)}(t)}{N_p^{(i)}(t)} - \beta_e^{(i)}E^{(i)}(t),$$

$$A^{(i)}(t+1) = A^{(i)}(t) + k_A\beta_e^{(i)}E^{(i)}(t) - \beta_{ac}^{(i)}A^{(i)}(t) - \gamma_{ar}^{(i)}A^{(i)}(t),$$

$$U^{(i)}(t+1) = U^{(i)}(t) + (1-k_A)\beta_e^{(i)}E^{(i)}(t) - \beta_{uc}^{(i)}U^{(i)}(t) - \gamma_{ur}^{(i)}U^{(i)}(t) - \gamma_{ud}^{(i)}U^{(i)}(t),$$

$$C^{(i)}(t+1) = C^{(i)}(t) + \beta_{uc}^{(i)}U^{(i)}(t) + \beta_{ac}^{(i)}A^{(i)}(t) - \gamma_{cr}^{(i)}C^{(i)}(t) - \gamma_{cd}^{(i)}C^{(i)}(t),$$

$$R_a^{(i)}(t+1) = R_a^{(i)}(t) + \gamma_{ar}^{(i)}A^{(i)}(t) - \gamma_{rs}^{(i)}R_a^{(i)}(t),$$

$$R_u^{(i)}(t+1) = R_u^{(i)}(t) + \gamma_{ur}^{(i)}U^{(i)}(t) - \gamma_{rs}^{(i)}R_u^{(i)}(t),$$

$$D_u^{(i)}(t+1) = D_u^{(i)}(t) + \gamma_{ud}^{(i)}U^{(i)}(t),$$

$$R_c^{(i)}(t+1) = R_c^{(i)}(t) + \gamma_{cr}^{(i)}C^{(i)}(t) - \gamma_{rs}^{(i)}R_c^{(i)}(t),$$

$$D_c^{(i)}(t+1) = D_c^{(i)}(t) + \gamma_{cd}^{(i)}C^{(i)}(t),$$

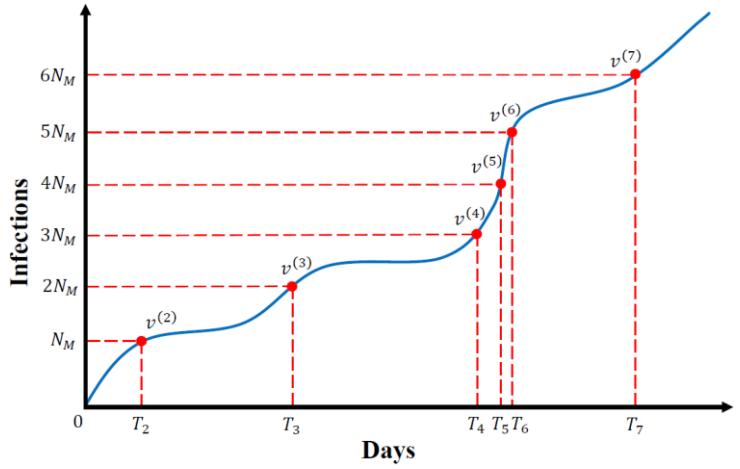
$$n_V^{(i)}(t) = \min\{n_{V,Thr}, f_v^s(t)\},$$

$$N_p^{(i)}(t+1) = (1+\gamma_p)N_p^{(i)}(t) - \gamma_{ud}^{(i)}U^{(i)}(t) - \gamma_{cd}^{(i)}C^{(i)}(t).$$

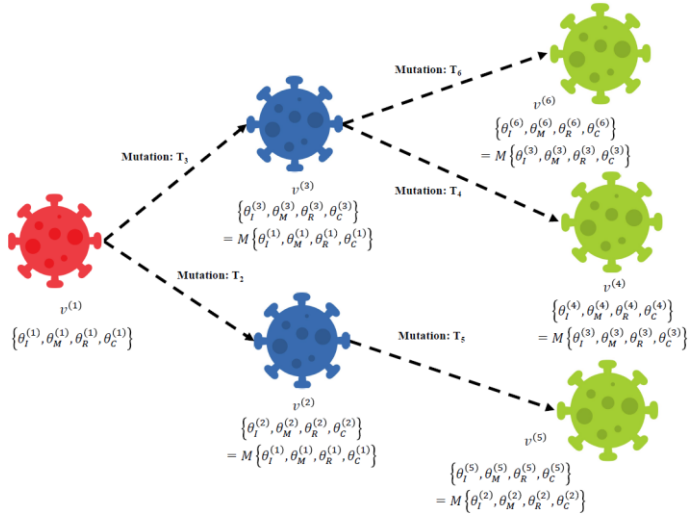
Methods: Variant mutation and vaccination simulation

Variant mutation

A mutation occurs when the number of infections reaches a threshold.

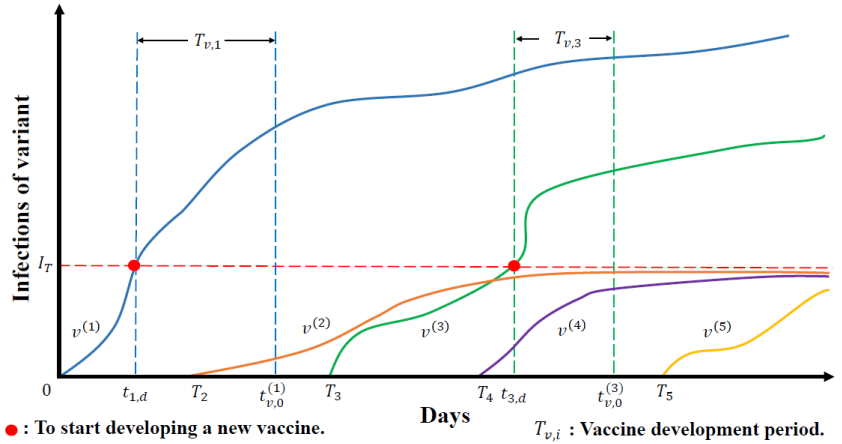


One variant will likely mutate multiple times to generate different new variants.

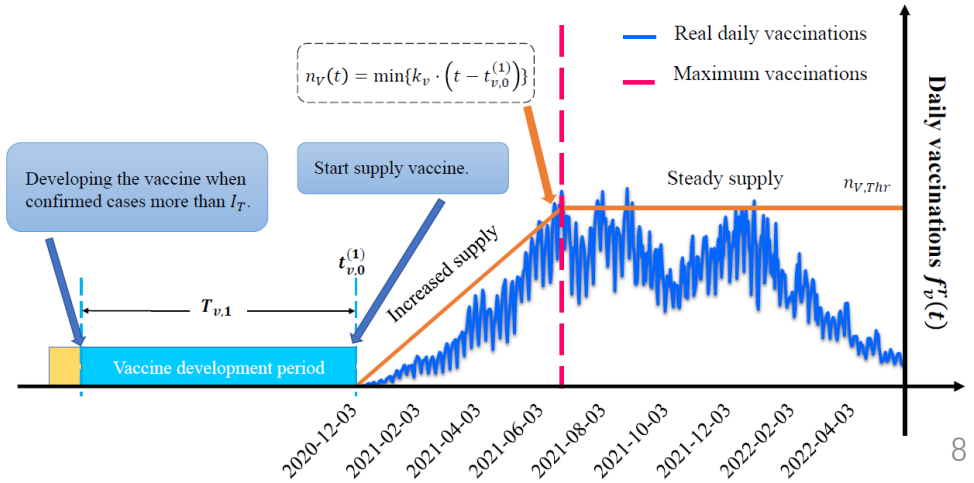


Vaccination

If a variant is confirmed as a "dangerous variant", then a vaccine will be developed against it.



In the early stages of the production of a new vaccine, the vaccine supplies are always insufficient.



Methods: Intervention simulation

Different parameters correspond to different interventions, and changing them can simulate changes in intervention intensity.

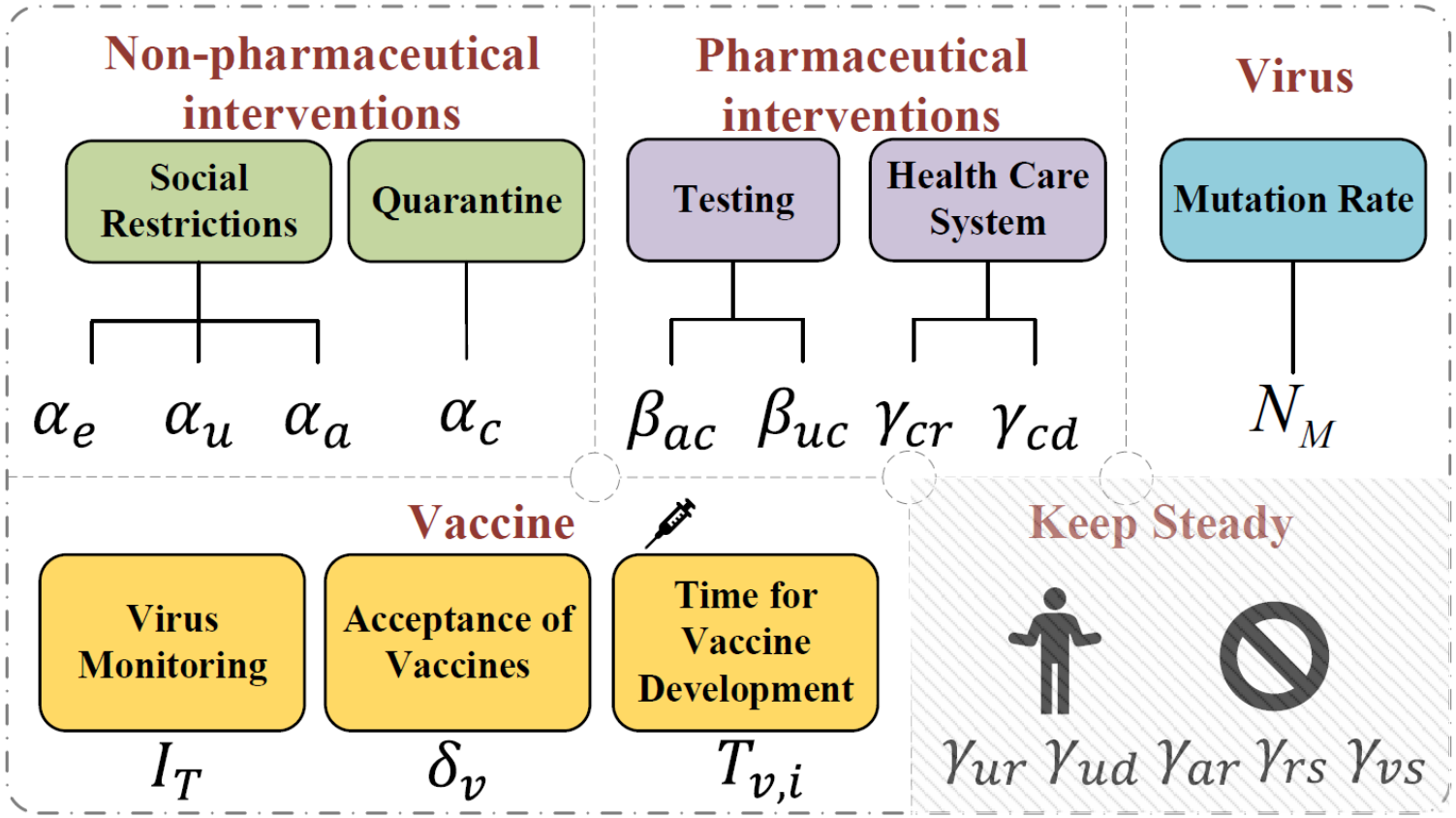
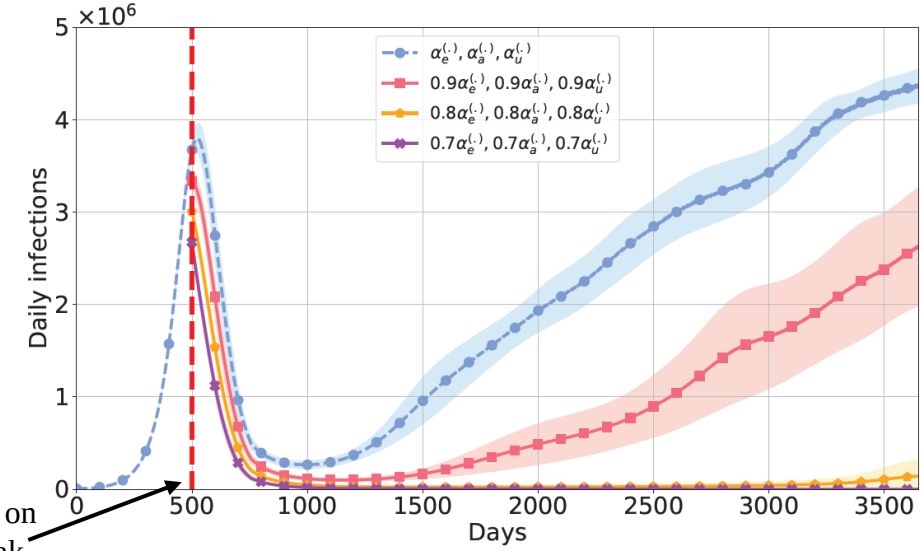


Fig. Factors can be varied by intervention measures in the framework.

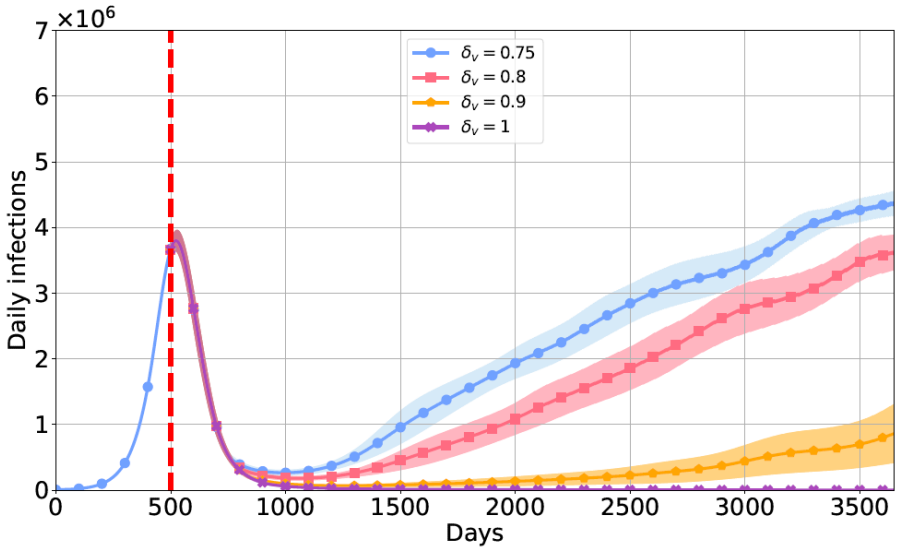
Experimental Results

Single intervention



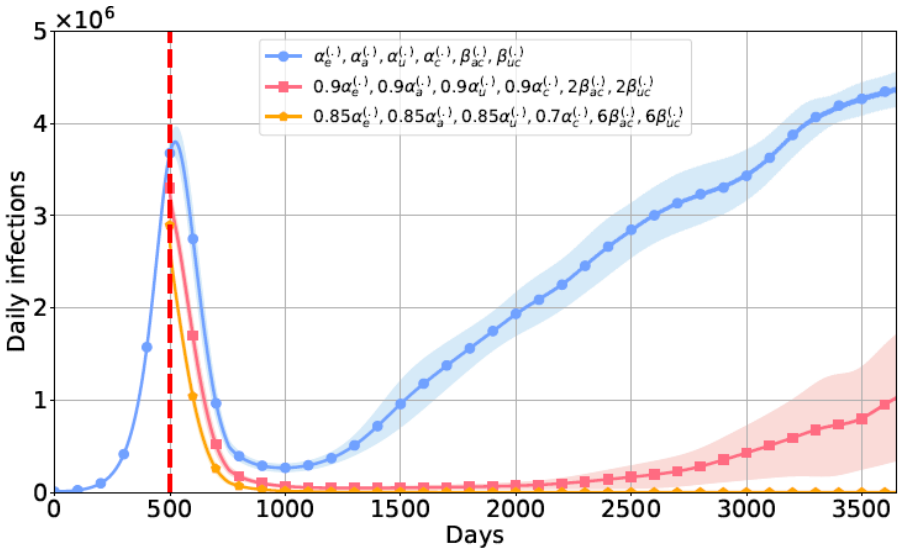
Interventions were performed on the 500th day after the outbreak.

Enhancing social restriction

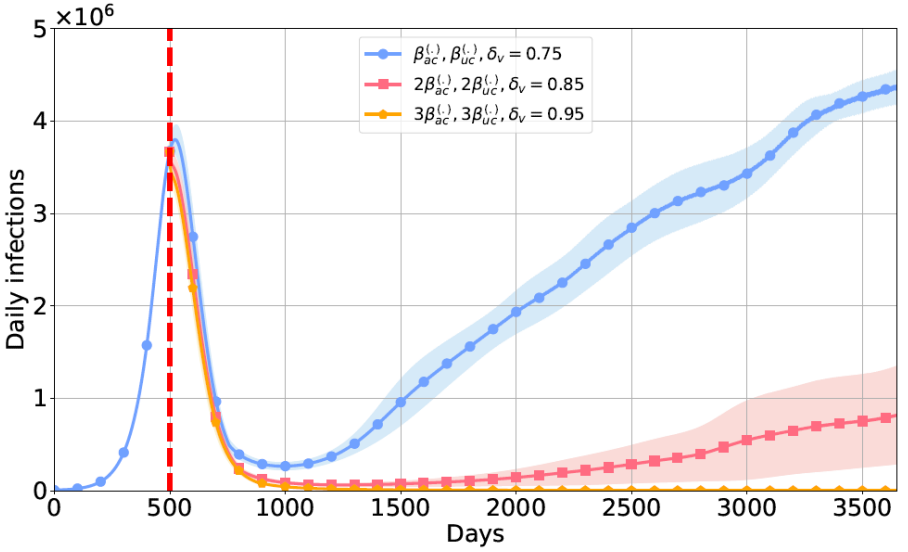


Increase vaccine acceptation rate

Combined intervention



Combined intervention strategy 1



Combined intervention strategy 3

Experimental Results

For single intervention, strictly restricting social contact and enhancing the vaccine acceptance rate yield the most optimal intervention for eradicating the pandemic.

Combined interventions are highly effective in containing or even eradicating the pandemic.

The influence of factors.

Single intervention	Intervention scenario	Almost no effect	Contain	Greatly contain	Eradicate	Criterion p_{it}	Criterion p_{it}^{D}
Non-pharmaceutical interventions							
(1) Restrict social contact	$0.72\alpha_e^{(\cdot)}, 0.72\alpha_u^{(\cdot)}, 0.72\alpha_a^{(\cdot)}$				✓	**0.00036**	0.05662
(2) Strict Quarantine of confirmed cases	$0\alpha_c^{(\cdot)}$			✓		0.38394	0.29073
Pharmaceutical interventions							
(3) Increased recovery rate	$9\gamma_{cr}^{(\cdot)}$		✓			0.61740	0.15102
(4) Reduce mortality rate	$0.1\gamma_{cd}^{(\cdot)}$	✓				0.95451	0.29633
(5) Enhance testing capacity	$9\beta_{ac}^{(\cdot)}, 9\beta_{uc}^{(\cdot)}$		✓			0.51148	0.32556
Vaccines							
(6) Enhance vaccine acceptance rate	$\gamma_v = 1$				✓	**0.00067**	0.08809
(7) Accelerating vaccine development	$T_{V,i} = 30$ days	✓				1.00000	1.00000
(8) Enhance virus monitoring	$I_T = 10,000$	✓				1.00000	0.98548
Natural process							
(9) Slowed virus mutation rate	$100N_M$			✓		0.22690	0.26870
Combination of intervention strategies							
(10) Combination strategy 1: (1) + (2) + (5)	$0.8\alpha_e^{(\cdot)}, 0.8\alpha_u^{(\cdot)}, 0.8\alpha_a^{(\cdot)}; 0.7\alpha_c^{(\cdot)}; 3\beta_{ac}^{(\cdot)}, 3\beta_{uc}^{(\cdot)}$				✓	**0.00006**	0.05035
(11) Combination strategy 2: (2) + (3) + (5)	$0\alpha_c^{(\cdot)}; 9\gamma_{cr}^{(\cdot)}; 9\beta_{ac}^{(\cdot)}, 9\beta_{uc}^{(\cdot)}$			✓		0.00243	0.02343
(12) Combination strategy 3: (5) + (6)	$3\beta_{ac}^{(\cdot)}, 3\beta_{uc}^{(\cdot)}; \gamma_v = 0.95$				✓	**0.00061**	0.07008

Note: Almost no effect: $0.9 \le p_{it}$; Contain: $0.4 \le p_{it} \le 0.9$; Greatly contain: $0.001 \le p_{it} \le 0.4$; Eradicate: $p_{it} \approx 0$ ($p_{it} \le 0.001$).

Criterion

$$p_{it} = \frac{\sum_{t=3551}^{3650} I_{all}^{(it)}(t)}{\sum_{t=3551}^{3650} I_{all}^{(0)}(t)}$$

$$p_{it}^{D} = \frac{\sum_{t=1}^{3650} D_{all}^{(it)}(t)}{\sum_{t=1}^{3650} D_{all}^{(0)}(t)}$$

Conclusion

- We proposed a novel epidemiological framework for simulating the emergence of realistic COVID-19 virus mutations and the transmission of multiple variants with human interventions.

- In scenarios with single interventions, strictly restricting social contact and enhancing the vaccine acceptance rate are the optimal interventions to contain and eradicate epidemics when considering the occurrence of virus mutations and multiple variants coexistence.

- By combining several intervention strategies, results show that the combined interventions are highly effective in containing or even eradicating the pandemic.

- Without any intervention, multiple variants can lead to multiple outbreaks.

Contents

Identifying Epidemic Transmission based on Metaheuristic and Dynamic Systems

Research Background and Significance

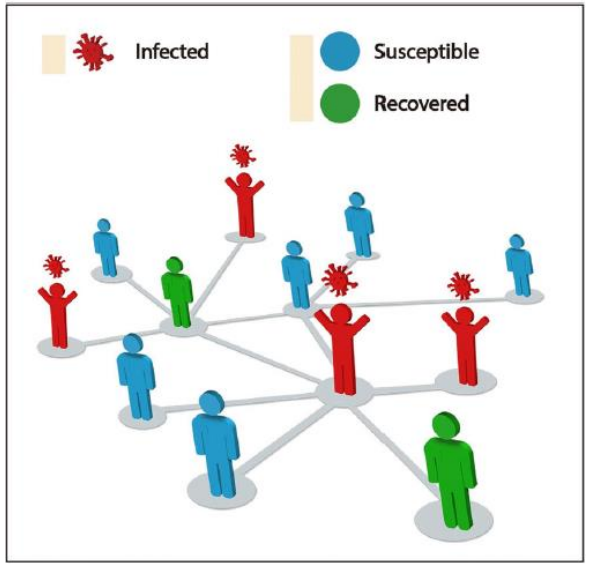


Fig. Contact network with susceptible, infected, and recovered individuals.

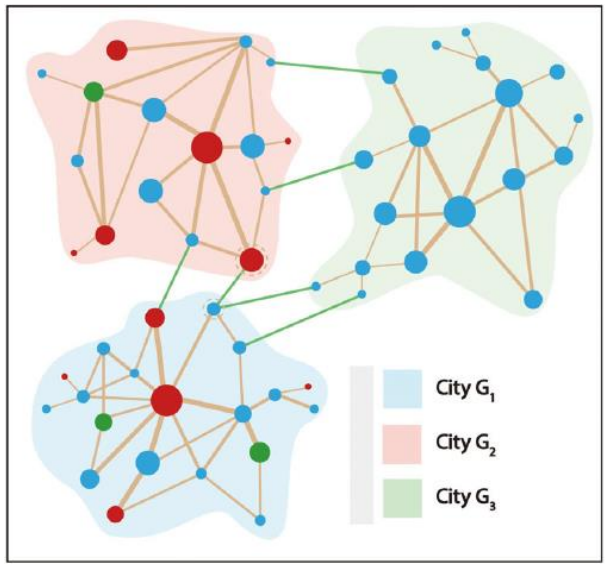


Fig. Contact network with 3 sub-networks (cities).

Motivation: Migration of individuals between intercity; A lot of unknown parameters are hard to determine.

Research aim: Predict the dynamic change in the number of COVID-19 infections based on the epidemic model and the intercity human migration network.

Achievement:

- Zhan C, **Zheng Y**, Lai Z, Hao T, Li B. Identifying epidemic spreading dynamics of COVID-19 by pseudocoevolutionary simulated annealing optimizers[J]. Neural Computing and Applications, 2020: 1-14. **(JCR Q2)**

Methods: SEIR-Migration

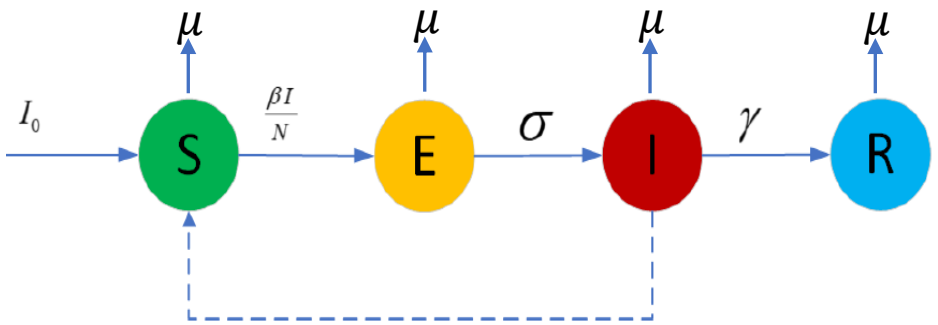


Fig. Susceptible-Exposed-Infected-Recovered model (SEIR). Susceptible (S), exposed (E), infected (I), and recovered/removed (R).

Mathematical definition of SEIR:

$$\Delta S(t) = -\frac{\beta S(t) I(t)}{N} + \mu(N - s(t))$$

$$\Delta E(t) = \frac{\beta S(t) I(t)}{N} - (u + \sigma) E(t)$$

$$\Delta I(t) = \sigma E(t) - (u + \gamma) I(t)$$

$$\Delta R(t) = \gamma I(t) - \mu R(t)$$

$$N = S(t) + E(t) + I(t) + R(t)$$

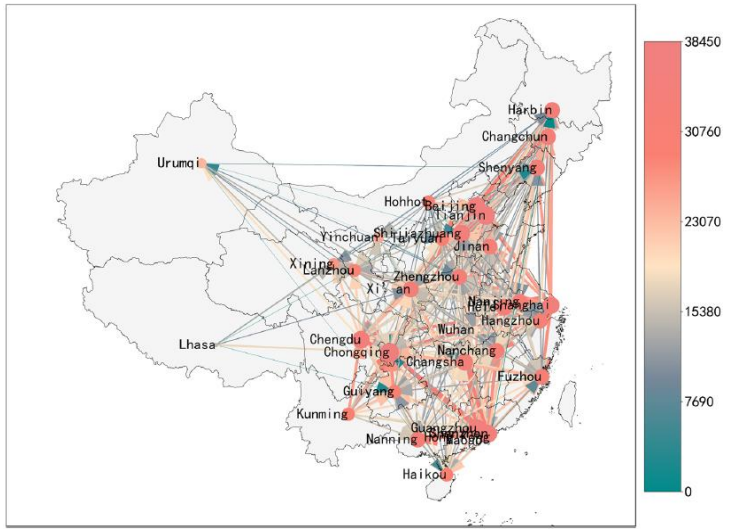


Fig. Intercity travel network of main cities in China on February 10, 2020 with 367 cities.

Migration matrix based on Baidu migration data:

$$M(t) = \begin{bmatrix} m_{11}(t) & m_{12}(t) & \cdots & m_{1N}(t) \\ m_{21}(t) & m_{22}(t) & \cdots & m_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1}(t) & m_{N2}(t) & \cdots & m_{NN}(t) \end{bmatrix}$$

Methods: SEIR-Migration

Mathematical definition of SEIR-Migration:

Each equation represents one type of individual transmission.

$$\Delta l_j(t) = \kappa_j E_i(t) - \gamma_j l_j(t) + k_I \left(\sum_{i=1}^{N} \left(\frac{l_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{l_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right)$$

$$\Delta E_j(t) = \frac{\beta_j}{N_j^s(t)} l_j(t) S_j(t) + \frac{\alpha_j}{N_j^s(t)} E_j(t) S_j(t) - \kappa_j E_i(t) + \sum_{i=1}^{N} \left(\frac{E_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{E_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}$$

$$\Delta S_j(t) = -\frac{\beta_j}{N_j^s(t)} I_j(t) S_j(t) - \frac{\alpha_j}{N_j^s(t)} E_j(t) S_j(t) + \sum_{i=1}^{N} \left(\frac{S_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{S_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}$$

$$\Delta R_j(t) = \gamma_j I_j(t)$$

$$\Delta P_j(t) = \sum_{i=1}^{N} m_{ij}(t) - \sum_{i=1}^{N} m_{ji}(t),$$

$$\Delta N_j^s(t) = k_l \left(\sum_{i=1}^{N} \left(\frac{l_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{l_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right) + \sum_{i=1}^{N} \left(\frac{E_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{E_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}$$
$$+ \sum_{i=1}^{N} \left(\frac{S_i(t) m_{ij}(t)}{P_i(t)} \right) - \frac{S_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)},$$

Methods: Pesudocevoluionary simulated annealing optimizers

1. Tune all the $5K+3$ parameters in the main procedure. Then, the root mean square percentage error (RMSPE) to measure the difference between the real infection and the estimated infection.

2. The estimated infected individuals generated by this extended SEIR-migration model.

3. Find the index of the $M1$ largest RMSPE and only tune the parameter sets of the corresponding cities.

4. Randomly select $M2$ cities and adjust their parameters to avoid the parameters of some cities that have not been adjusted.

Algorithm 3 Pesudocoevolutionary algorithm for estimating the optimal parameter set Θ^*

Input: The set of unknown parameters and initial number of infected and exposed individuals of each city $\Theta = \{I_{WH,0}, E_{WH,0}, k_I, \theta_1, \theta_2, \cdots, \theta_K\}$, where $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}$ and $j = 1, 2, \cdots K$;

Output: Optimal parameter set Θ^*;

Initialisation : Initialize temperature T, and random starting point

$$\Theta_0 = \Theta_L + k_{rand} * (\Theta_U - \Theta_L),$$

The index of adopted cities is $\Phi = 1, 2, \cdots, K$.

Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

LOOP Process

for $i = 0$ to M **do**

  for $j = 1$ to K **do**

    Set the model parameter as Θ_0 and apply Algorithm-1 to derive $\{\hat{I}(t_0|\Theta_0), \hat{I}(t_1|\Theta_0), \cdots, \hat{I}(t_N|\Theta_0)\}$.

    Using evaluation criteria (13) to derive RMSPE_j for each city.

    Find the index of the $M1$ largest RMSPE_j and set $\Phi = \{a_1, a_2, \cdots, a_{M1}\}$, where a_i represents the index of the i-th largest RMSPE_j.

    Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

    Randomly generate $M2$ unique integers a_i between 1 and K and set $\Phi = \{a_1, a_2, \cdots, a_{M2}\}$.

    Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

  end for

end for

$$\Theta^* \leftarrow \Theta_0$$

return Optimal parameter set Θ^*.

Experimental Results

The proposed model can estimate the daily number of infected, exposed, and recovered individuals in all 367 cities.

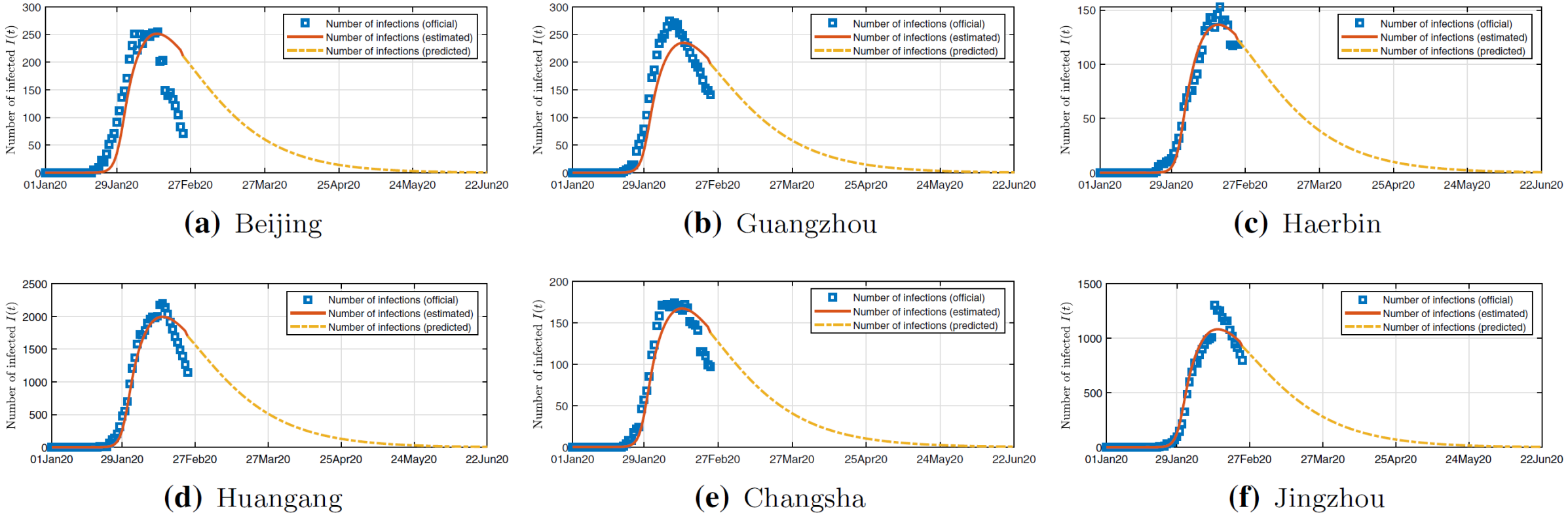


Fig. Estimated historical data and prediction of the number of infected individuals for next 150 days.

Conclusion

- We integrate daily intercity migration data and traditional SEIR model to develop an extended SEIR model.
- A novel pseudocoevolutionary simulated annealing algorithm is proposed and have a best performance by comparing the estimation result with simulated annealing, particle swarm optimization, and pattern search algorithms.
- The migration control may extremely effective in controlling the spread of the epidemic.

Contents

Forecasting Infectious Disease Transmission based on the Hybrid Machine Learning Model

Research Background and Significance

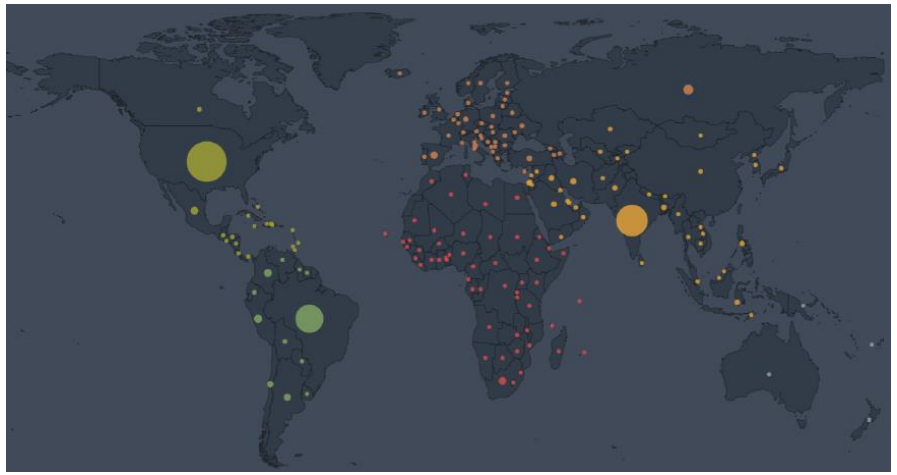


Fig. The cumulative confirmed cases in the world up to Sep 2020.

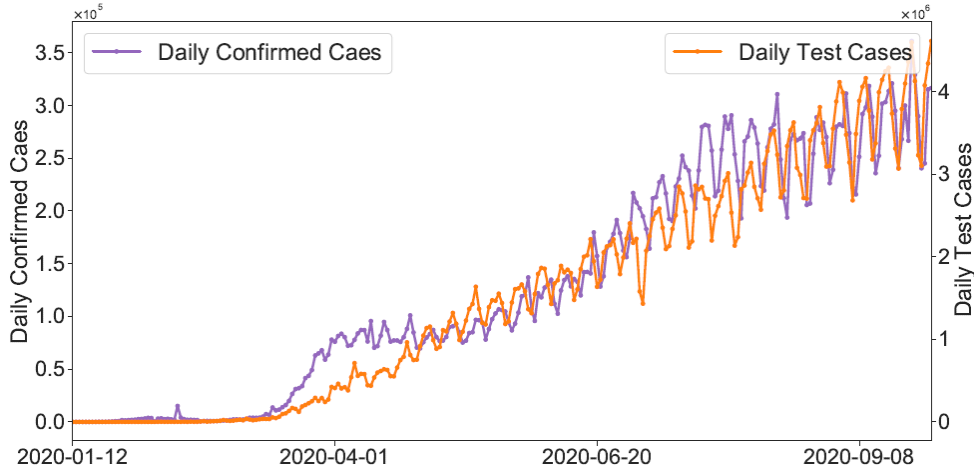


Fig. The daily confirmed cases and tests over the world up to Sep 30, 2020.

Motivation:

The daily COVID-19 cases in different areas are highly volatile and variable; Epidemic cases are affected by multiple factors

Research aim: Predicting the dynamic change in the number of COVID-19 infections and the number of nucleic acid tests worldwide.

Achievement:

- Zhan C, **Zheng Y**, Zhang H, Wen Q. Random-Forest-Bagging Broad Learning System with Applications for COVID-19 Pandemic[J]. IEEE Internet of Things Journal, 2021. **(JCR Q1)**
- Lin J, Tan M, **Zheng Y**, Wu K, Zhan C. "Detection Capability Prediction Based on Broad Learning System During The COVID-19 Pandemic." 2021 16th international conference on intelligent systems and knowledge engineering (ISKE). IEEE, 2021. **(EI)**

Methods: Broad Learning Systems (BLS), Bagging

1. The broad learning system employs incremental learning, expanding neural nodes and updating network weights as needed, without stacking layers.

2. Bagging is the ensemble learning method that is commonly used to reduce variance within a noisy data set.

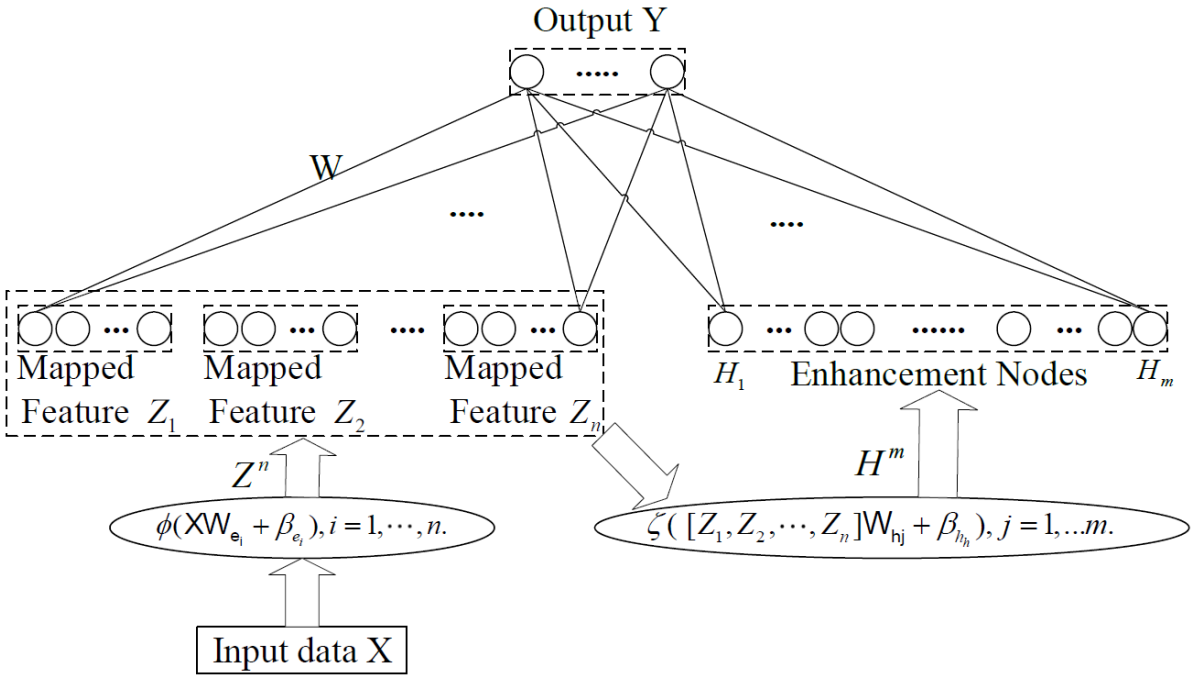


Fig. Structure of BLS.

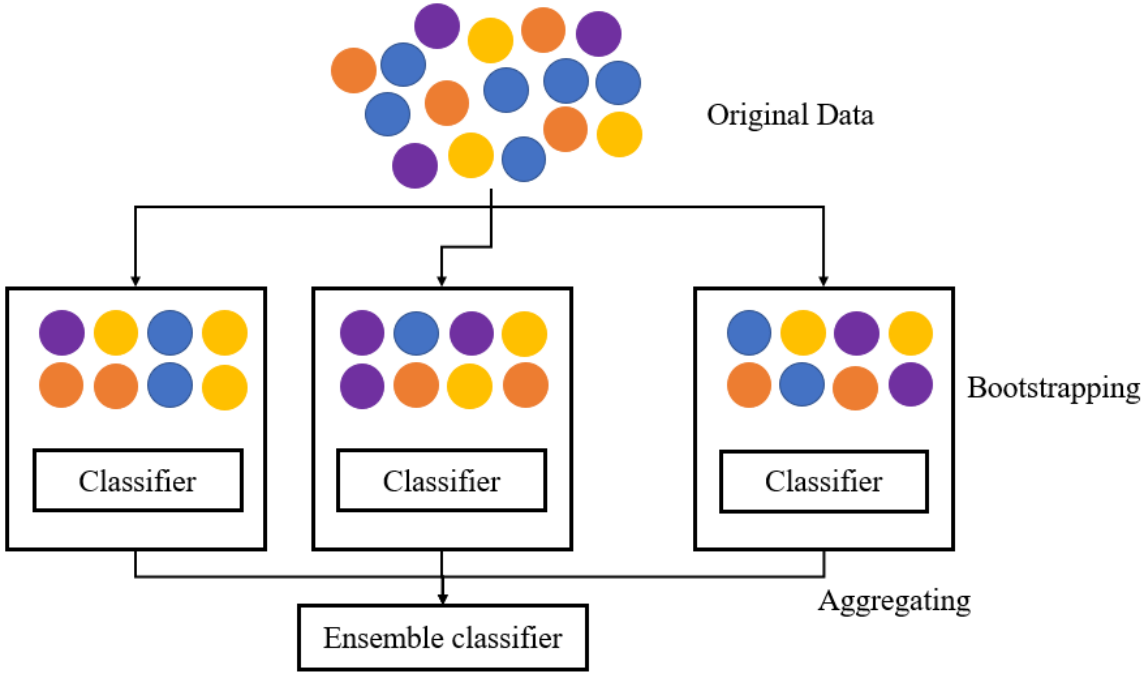


Fig. Structure of Bagging.

Methods: Random Forest-Bagging-BLS (RF-Bagging-BLS)

The Random Forest model and ensemble learning-bagging are adopted to enhance the performance of BLS for the prediction of the spread of COVID-19.

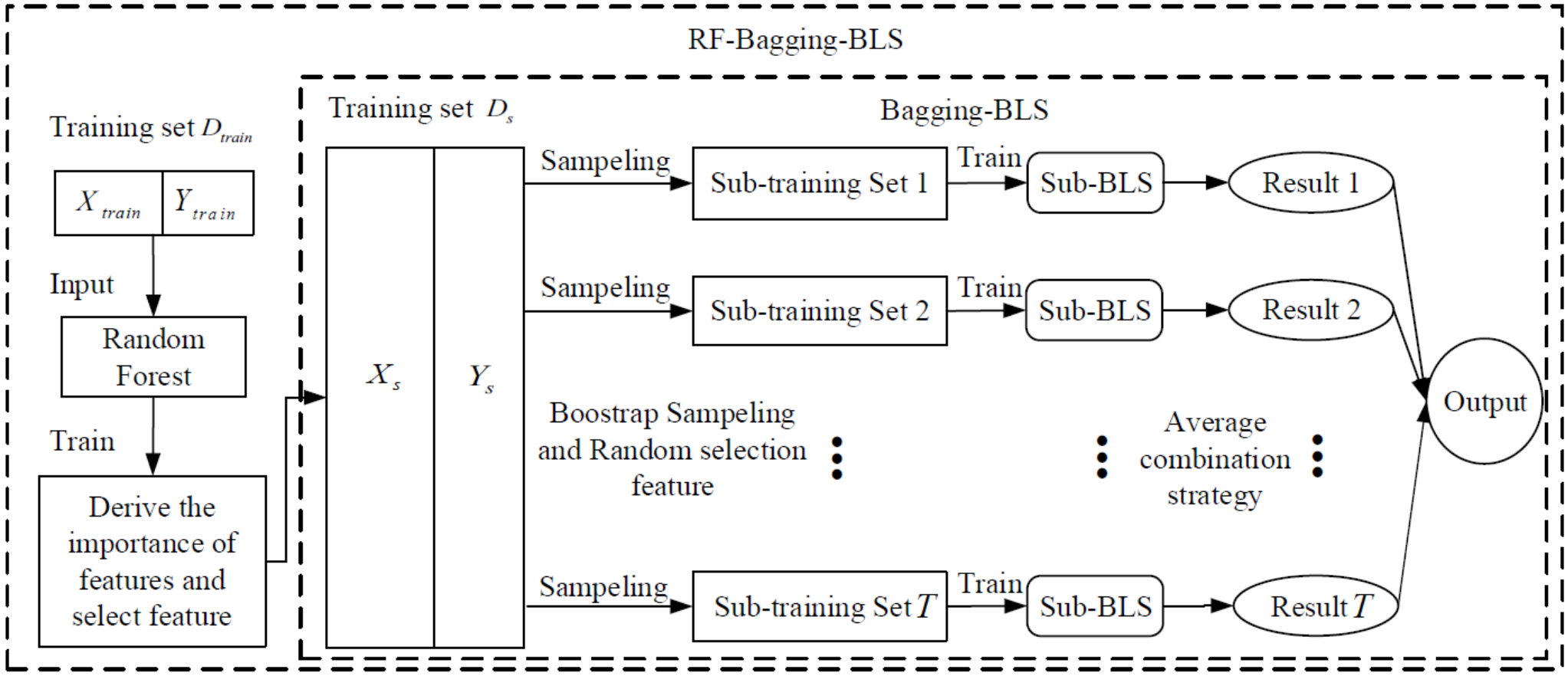


Fig. Structure of RF-Bagging-BLS.

Experimental Results

Establish data set on the COVID-19 pandemic in 184 countries and 1241 areas from Dec 2019 to Sep 2020.

For comparison, the correlation analysis was used in feature selection for other classical models.

Tab. Feature after feature engineering

	Original features
$I_C(t)$	Cumulative confirmed cases at time t
$R(t)$	Total recovered cases at time t
$D(t)$	Death toll at time t
$N_T(t)$	Cumulative COVID-19 tests at time t
x_{LA}	The latitude of the geographical location of an area
x_{LO}	The longitude of the geographical location of an area
x_{CT}	The continent to which an area belongs
x_D	The level of economic development of an area
x_P	The population of an area
	Augmented features
$\Delta I_C(t)$	Daily confirmed cases
$\Delta R(t)$	Daily recovered cases
$\Delta D(t)$	Daily deaths
$I(t)$	Active COVID-19 patients
$\Delta N_T(t)$	Daily COVID-19 tests
$r_I(t)$	Daily growth rate of confirmed cases
$r_R(t)$	Daily growth rate of daily recover cases
$r_D(t)$	Daily growth rate of daily death cases

Tab. Selected features in the classical models

	Selected features	Description
Original features	$I_C(t-m)$	Cumulative confirmed cases at day $t-m$ $(m=1,\cdots,7)$
	$R(t-m)$	Total recovered cases at day $t-m$ $(m=1,\cdots,7)$
	$D(t-k)$	Death toll at day $t-m$ $(m=1,\cdots,7)$
	$N_T(t-m)$	Cumulative COVID-19 tests at day $t-m$ $(m=1,\cdots,7)$
	x_P	The population of an area
Augmented features	$I(t-m)$	Activate COVID-19 patients at day $t-m$ $(m=1,\cdots,7)$
	$\Delta I_C(t-m)$	Daily confirmed cases at day $t-m$ $(m=1,\cdots,7)$
	$\Delta N_T(t-m)$	Daily COVID-19 tests at day $t-m$ $(m=1,\cdots,7)$

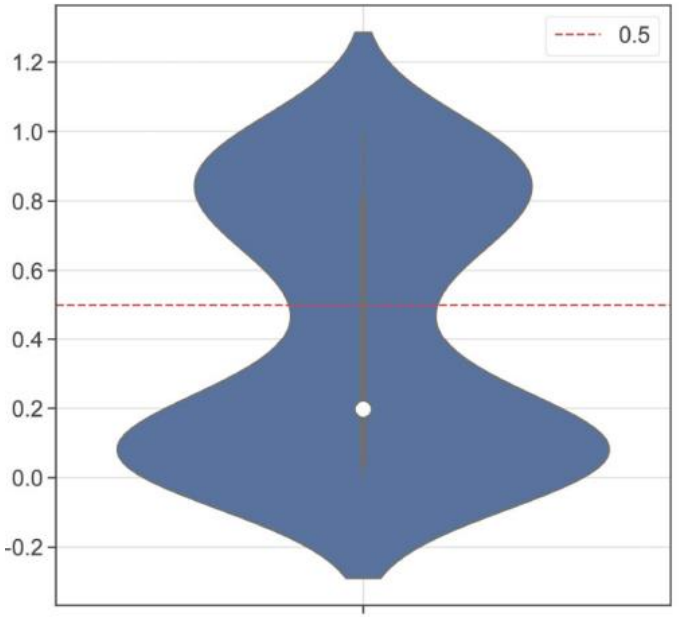
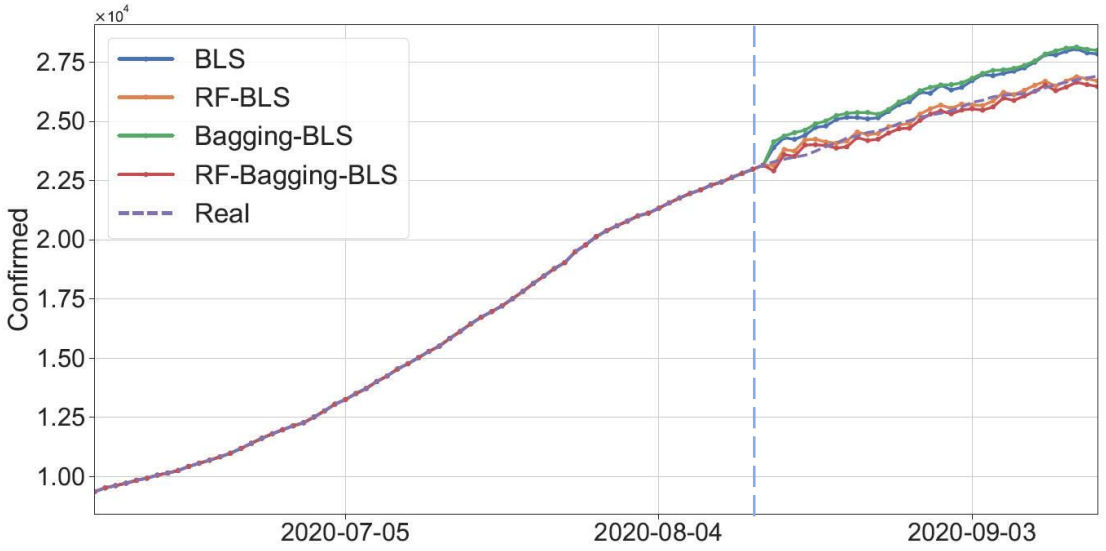
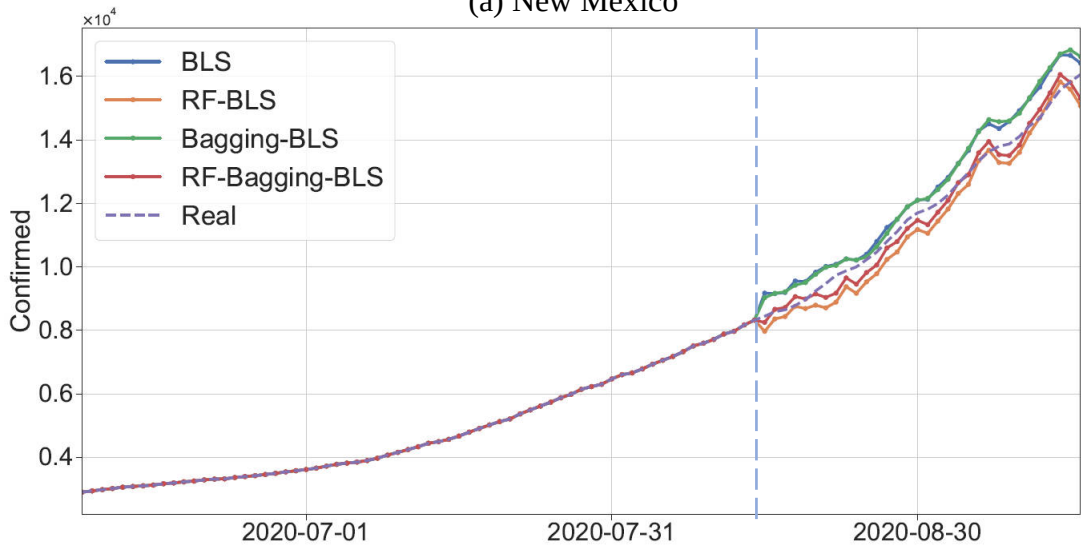


Fig. Distribution of correlation coefficients between features and output.

Experimental Results



(a) New Mexico



(b) North Dakota

Fig. Predictive results of different areas.

Experimental results show that the proposed RF-Bagging-BLS achieve the best value among all the comparison algorithms.

We also test CNN, LSTM, and GRU based on our data set. However, these methods are easily overfitting.

Tab. Prediction result of $y(t+n)$

Method	RMSE	MAE	R^2	MAD	R^2_{adj}	MAPE
LR	3394.6494	1601.8012	0.9994	642.4868	0.9994	3.2467
KNN	32267.9550	16772.8781	0.9483	6280.7837	0.9483	22.3130
DT	24733.7204	10304.4542	0.9706	1423.0000	0.9696	10.8748
SVR	53319.4198	23591.0204	0.8634	8708.3930	0.8589	63.6835
Ada	17861.0283	5933.6822	0.9846	1196.0000	0.9842	5.5883
RF	23770.2149	8554.5544	0.9728	1218.9372	0.9719	6.7004
GBDT	21103.7782	7325.3585	0.9786	977.8010	0.9778	7.2672
ET	19635.1575	6997.1890	0.9814	1469.4092	0.9808	6.7076
CAT	25744.2812	8687.9153	0.9681	1351.8562	0.9671	7.3150
LGB	25470.3411	9120.9554	0.9688	973.4710	0.9678	7.9324
XGB	24354.7123	8455.0691	0.9715	1082.3067	0.9705	6.8996
BLS	2269.8906	1362.8407	0.9997	766.4345	0.9997	6.7834
RF-BLS	2042.4723	1018.2183	0.9997	565.5706	0.9997	6.4979
Bagging-BLS	2872.1029	1475.1688	0.9996	697.3510	0.9995	3.9903
RF-Bagging-BLS	**1989.1970**	**952.5739**	**0.9998**	**432.3244**	**0.9998**	**3.0090**

Experimental Results

Previous work:
In some cases, data-driven methods perform well when the output data are close to a uniform or normal distribution.

However, the predictive value with a uniform or normal distribution does not help in improving the predictive performance in this study.

Tab. Predictive result of $log_{10}\ y(t+n)$

Method	RMSE	MAE	R^2	MAD	R^2_{adj}	MAPE
LR	57387.6106e+9	76255.7453e+8	-1.5816e+17	21649.1423	-1.6344e+17	2.5650+9
KNN	32589.3837	13272.0041	0.9498	3875.6634	0.9472	13.7311
DT	20441.1788	9758.9613	0.9799	3664.0001	0.9792	31.8415
SVR	119751.7181	43284.9216	0.3113	6676.6545	0.2883	28.2169
Ada	27126.1988	9364.6234	0.9646	1033.5000	0.9634	5.9561
RF	23413.5558	8136.2741	0.9736	1090.4159	0.9727	5.2950
GBDT	25739.2592	9643.7920	0.9681	1791.4312	0.9671	8.3891
ET	21978.5323	**7612.5017**	0.9768	1292.7426	0.9760	6.1060
CAT	31788.7347	11952.3085	0.9515	1533.8260	0.9498	8.4482
LGB	24249.7400	8033.8196	0.9717	**939.8207**	0.9708	**5.4479**
XGB	29203.4126	9389.2712	0.9590	1107.6260	0.9576	5.5478
BLS	35079.3674	16463.1951	0.9409	9145.3791	0.9389	27.6222
RF-BLS	39131.9767	23896.6042	0.9265	9328.3896	0.9255	40.3023
Bagging-BLS	**13539.3023**	9882.3092	**0.9911**	7690.6102	**0.9909**	23.0960
RF-Bagging-BLS	35793.3709	19923.3754	0.9384	8895.5277	0.9376	39.5285

Tab. Predictive result of $y(t+n)$

Method	RMSE	MAE	R^2	MAD	R^2_{adj}	MAPE
LR	3394.6494	1601.8012	0.9994	642.4868	0.9994	3.2467
KNN	32267.9550	16772.8781	0.9483	6280.7837	0.9483	22.3130
DT	24733.7204	10304.4542	0.9706	1423.0000	0.9696	10.8748
SVR	53319.4198	23591.0204	0.8634	8708.3930	0.8589	63.6835
Ada	17861.0283	5933.6822	0.9846	1196.0000	0.9842	5.5883
RF	23770.2149	8554.5544	0.9728	1218.9372	0.9719	6.7004
GBDT	21103.7782	7325.3585	0.9786	977.8010	0.9778	7.2672
ET	19635.1575	6997.1890	0.9814	1469.4092	0.9808	6.7076
CAT	25744.2812	8687.9153	0.9681	1351.8562	0.9671	7.3150
LGB	25470.3411	9120.9554	0.9688	973.4710	0.9678	7.9324
XGB	24354.7123	8455.0691	0.9715	1082.3067	0.9705	6.8996
BLS	2269.8906	1362.8407	0.9997	766.4345	0.9997	6.7834
RF-BLS	2042.4723	1018.2183	0.9997	565.5706	0.9997	6.4979
Bagging-BLS	2872.1029	1475.1688	0.9996	697.3510	0.9995	3.9903
RF-Bagging-BLS	**1989.1970**	**952.5739**	**0.9998**	**432.3244**	**0.9998**	**3.0090**

Conclusion

- We collect and unified one comprehensive data set, including the epidemic spread data, geographic information, economic information, population, and COVID-19 testing information of 184 countries and 1241 areas.

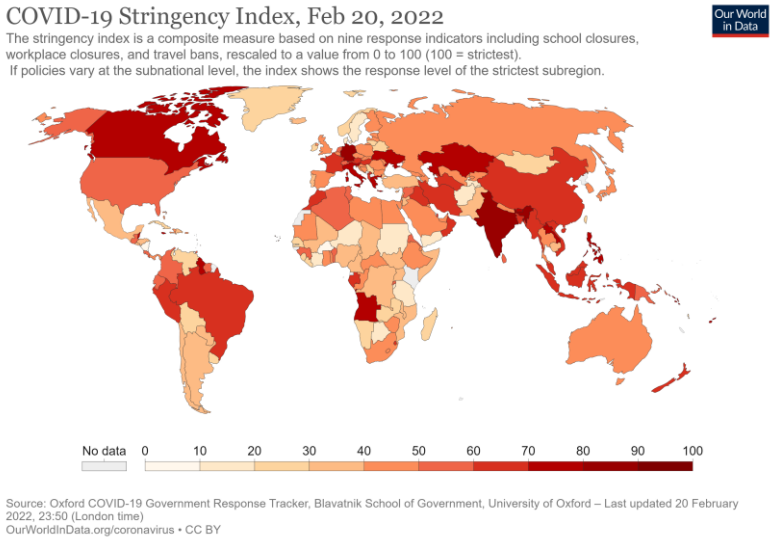
- A hybrid machine learning model of RF-Bagging-BLS is developed for predicting the COVID-19 trend in different countries, which showed promising performance in timely short-term forecasts without the requirement of epidemiological models.

- The predictive value with a uniform or normal distribution does not help in improving the predictive performance in the COVID-19 infection prediction.

Contents

Quantitation of Pandemic Outbreak Impact based on Statistical Model

Research Background and Significance



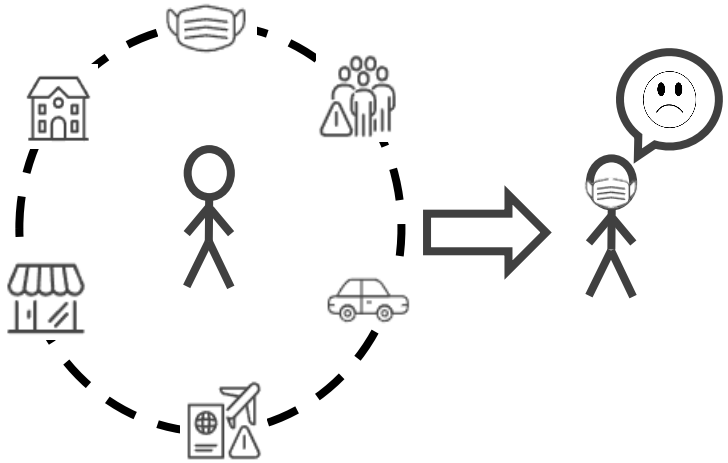


Fig. Global heat map of COVID-19 stringency index, February 20, 2022.

Motivation: The impact of the epidemic on the entertainment industry was unknown; no prediction tool for online game stream

Research aim: Quantifying the impact of the COVID-19 pandemic on movies and online games, and building the prediction models of box office and online player's numbers.

Achievement:

- **Zheng Y**, Zhen Q, Tan M, Hu H, Zhan C. "COVID-19's impact on the Box office: Machine Learning and Difference-in-Difference." 2021 16th international conference on intelligent systems and knowledge engineering (ISKE). IEEE, 2021. **(EI)**
- Wu S, Hu H, **Zheng Y**, Zhen Q, Zhang S, Zhan C. "The impact of COVID-19 on online games: Machine learning and Difference-In-Difference." CCF Conference on Computer Supported Cooperative Work and Social Computing. Springer, Singapore, 2021. **(EI)**

Methods: difference-in-difference (DID) model

Quantified the impact of the pandemic on daily box office and online players by the difference-in-difference method.

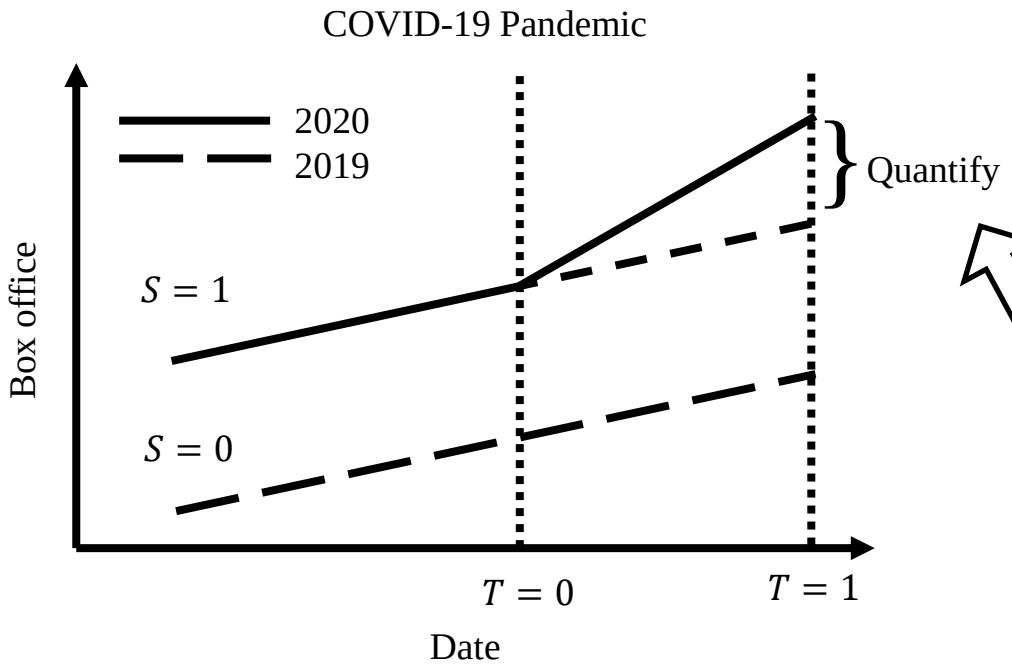


Fig. difference-in-difference model

Mathematical definition of regression analysis of DID:

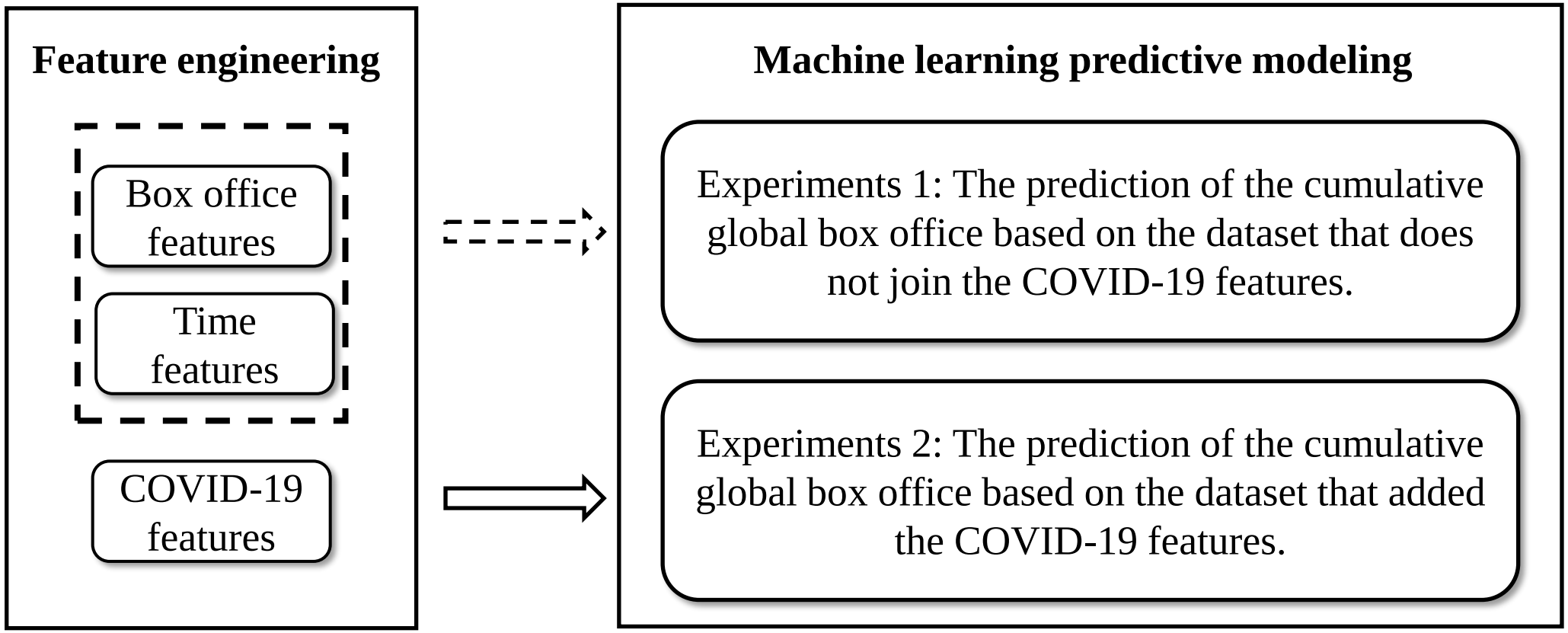
$$y = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 (T \cdot S) + \varepsilon$$

Calculation of intervention effect coefficient β_3:

$$\hat{\beta}_0 = \hat{E}(y \mid T = 0, S = 0)$$
$$\hat{\beta}_1 = \hat{E}(y \mid T = 1, S = 0) - \hat{E}(y \mid T = 0, S = 0)$$
$$\hat{\beta}_2 = \hat{E}(y \mid T = 0, S = 1) - \hat{E}(y \mid T = 0, S = 0)$$
$$\hat{\beta}_3 = [\hat{E}(y \mid T = 1, S = 1) - \hat{E}(y \mid T = 0, S = 1)] - [\hat{E}(y \mid T = 1, S = 0) - \hat{E}(y \mid T = 0, S = 0)]$$

Prediction Framework

Explored the impact of epidemic features on predictive modelling.



Experimental Results

Tab. Quantitative result.

Variable	Coefficient
$T \cdot S$	$-8.456^{***}(0.323)$
T	$1.876^{***}(0.151)$
S	$1.260^{***}(03157)$
ε	$14.23^{***}(0.120)$
R^2	0.590

Note : Robust standard errors in parentheses

$***p < 0.01, **p < 0.05, ^{*}p < 0.1$

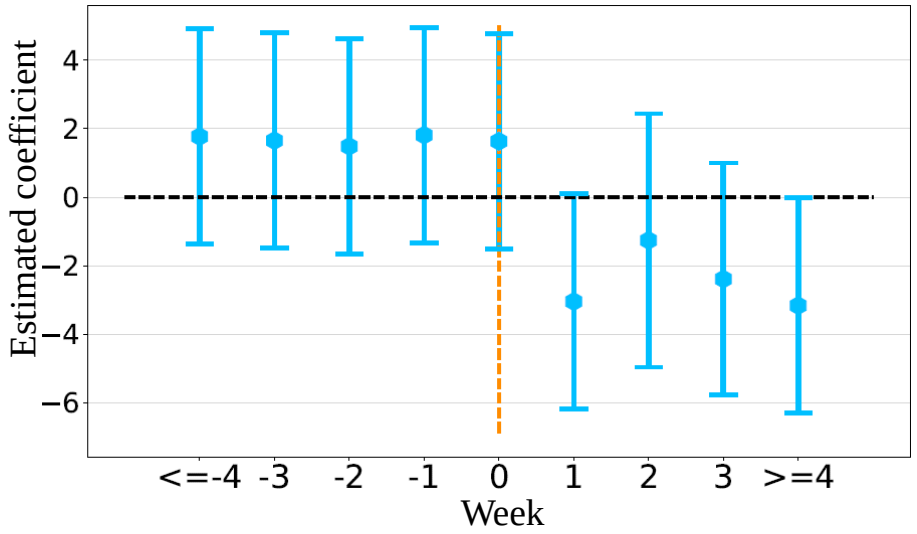


Fig. Parallel trend test result.

Tab. Prediction result of experiment without epidemic features.

Model	MAPE	RMSE	MAE	R^2
LR	10.6939	19232852.6366	13180724.7015	0.9277
GBDT	14.404	27070929.4065	21926058.8144	0.8569
Ada	30.5505	36755333.2194	30258660.7305	0.7361
RF	24.3341	37473895.2314	26794616.6136	0.7257
SVR	19.9977	38720566.4797	30448341.6686	0.7071
EXT	39.8577	75465184.4287	63664650.7227	−0.1124
DT	47.9107	93044797.2858	75406330.7113	−0.691
KNN	71.9418	139174211.146	118802756.3855	−2.7834
MLP	84.4943	154523408.4404	135487709.9276	−3.6639
LGB	151.9525	200942540.9506	187382988.0388	−6.8869

Tab. Prediction result of experiment with epidemic features.

Model	MAPE	RMSE	MAE	R^2
EXT	13.0756	18978870.5778	14320150.7542	0.9296
GBDT	16.8319	26273880.9723	19550281.2743	0.8652
MLP	14.8964	33027102.6926	24329214.3868	0.7869
DT	29.2268	53179893.8243	41400820.8479	0.4476
LR	36.3958	55167314.7062	48379764.724	0.4055
RF	36.046	69380517.3971	55358251.1509	0.0598
Ada	56.1211	74741304.749	65278153.5556	−0.0911
LGB	78.4685	79813783.3509	63971288.3595	−0.2443
SVR	89.9288	97213514.2064	86030594.5032	−0.8459
KNN	68.4771	136564744.1887	116181101.6124	−2.6428

Conclusion

- Established a research dataset including the information of movies, popular games, and COVID-19.
- Developed the Difference-in-Difference method to quantify the impact of the COVID-19 pandemic on the daily global box office and online game players. The results show that the pandemic has a negative effect on both box office and online game platforms.
- Used machine learning, ensemble learning, and deep learning to build a predicted model of daily global box office and online players. And explored the impact of the pandemic and human mobility on prediction models.

Contents

Association of Human Mobility and Weather on Mosquito Activity based on Statistical Model

Research Background and Significance

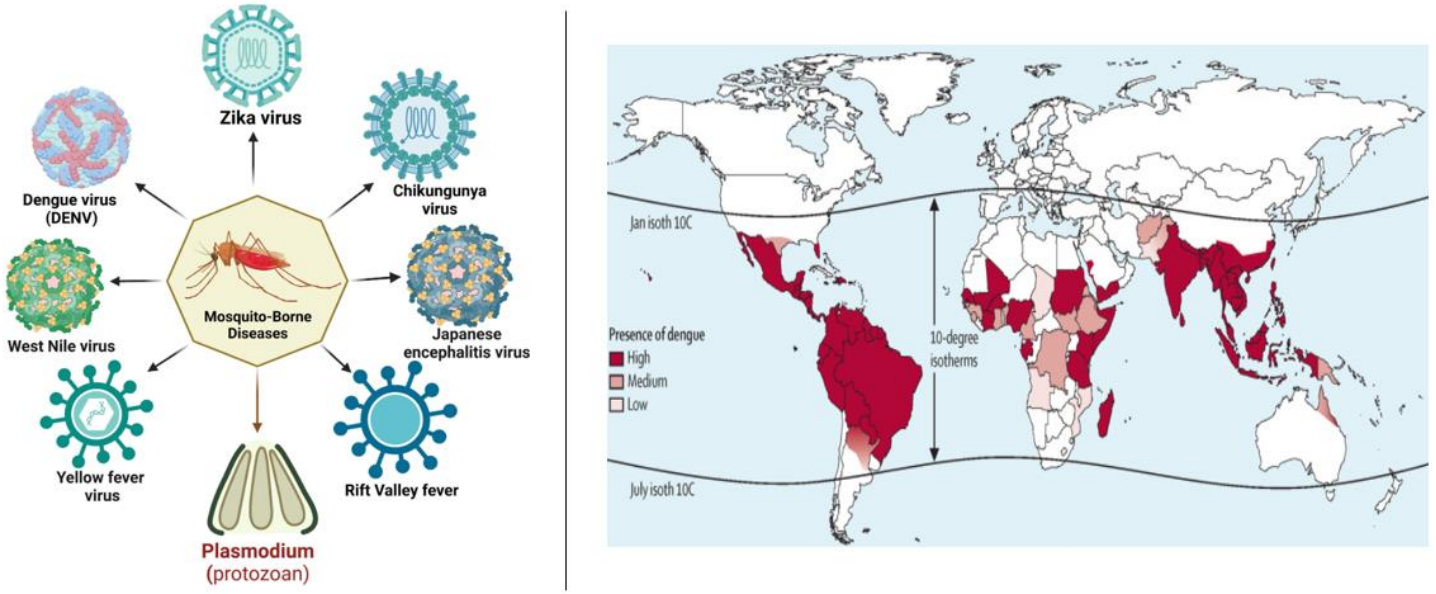


Fig. Mosquito-Borne-Diseases and transmission of Dengue Fever.

Motivation:

Human activities overlap with mosquito activities; human mobility and weather lag influences the abundance of mosquitoes is still unknown

Research aim: Assess the influence of human mobility and weather on the abundance of mosquito (*Aedes albopictus*).

Achievement:

- **Zheng Y**, Yue K, Wong, Eric W M, Yuan, H Y. Association between meteorological factors and human mobility with mosquito activity risk in Hong Kong. medRxiv (2024). (Submitted to journal)

Methods: Framework

Human mobility indices and weather conditions together with *Aedes albopictus* abundance and extensiveness were obtained. Distributed lag non-linear models with mixed-effects models were used to explore their influence in Hong Kong.

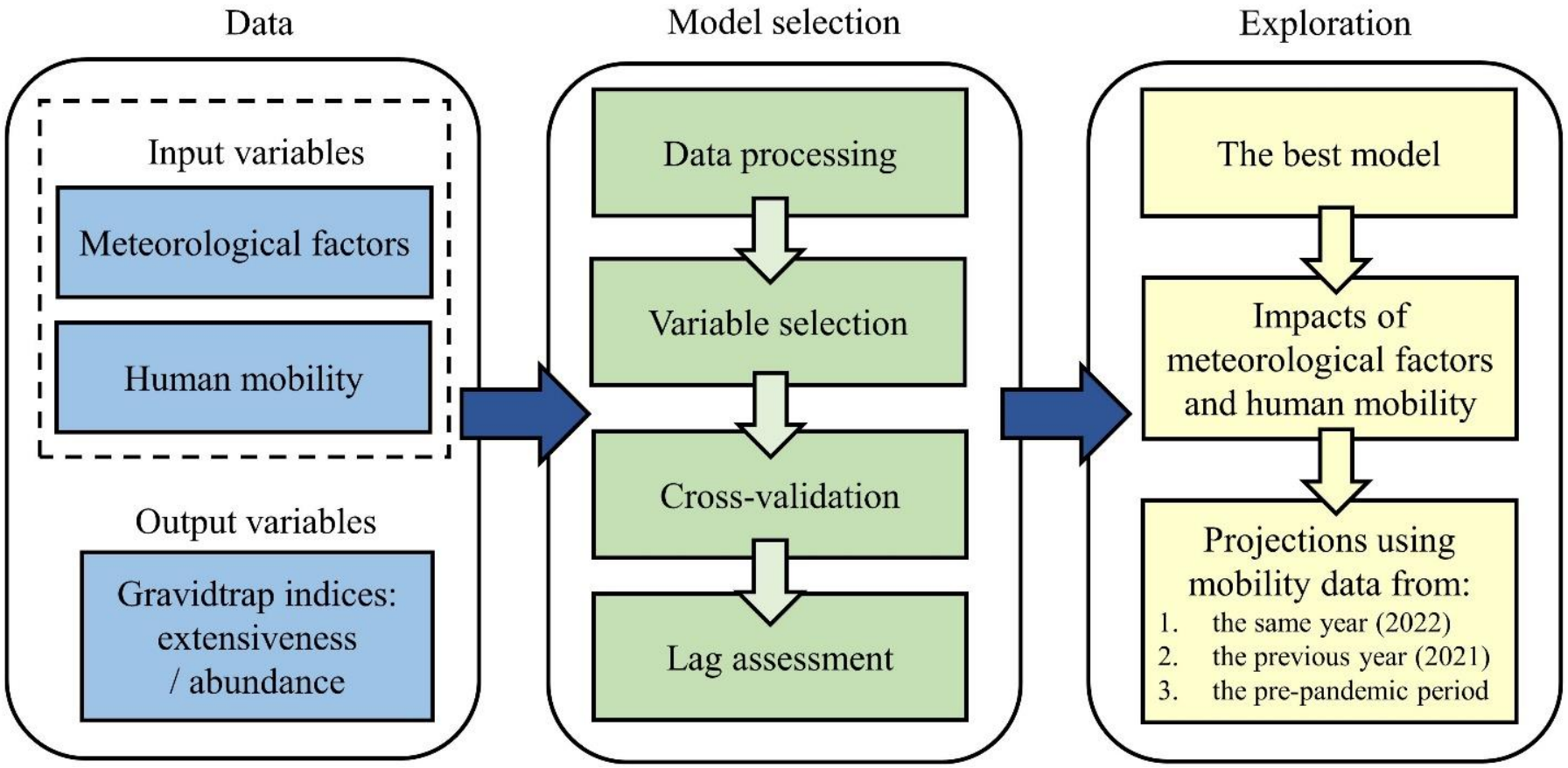


Fig. Research framework.

Methods: Framework and DLNM

Let λ_t be the mosquito number per 1000 traps at month t.

$$\lambda_t \sim NB(u_t, \kappa)$$

The regression model to predict mosquito abundance was:

$$\log(\lambda_t) = \beta + \gamma + S + f.w(T_t, l_1) + f.w(R_t, l_2) + m_{it} + \alpha$$

where $f.w(R_t, l_1)$ and $f.w(T_t, l_2)$ represent the nonlinear exposure-lag functions of total rainfall R_t from 0 to l_1 months and mean temperature T_t from 0 to l_2 months in t^{th} month, respectively; S is the area random effect; γ is the monthly random effect; β is the yearly random effect; α is the intercept; and m_t is the human mobility index in one category (e.g. residential areas, workplaces, or parks) at t^{th} month.

Experimental Results: Data set

Data period: from April 2020 to August 2022.

Daily data; 53 mosquito monitoring traps; 13 weather stations.

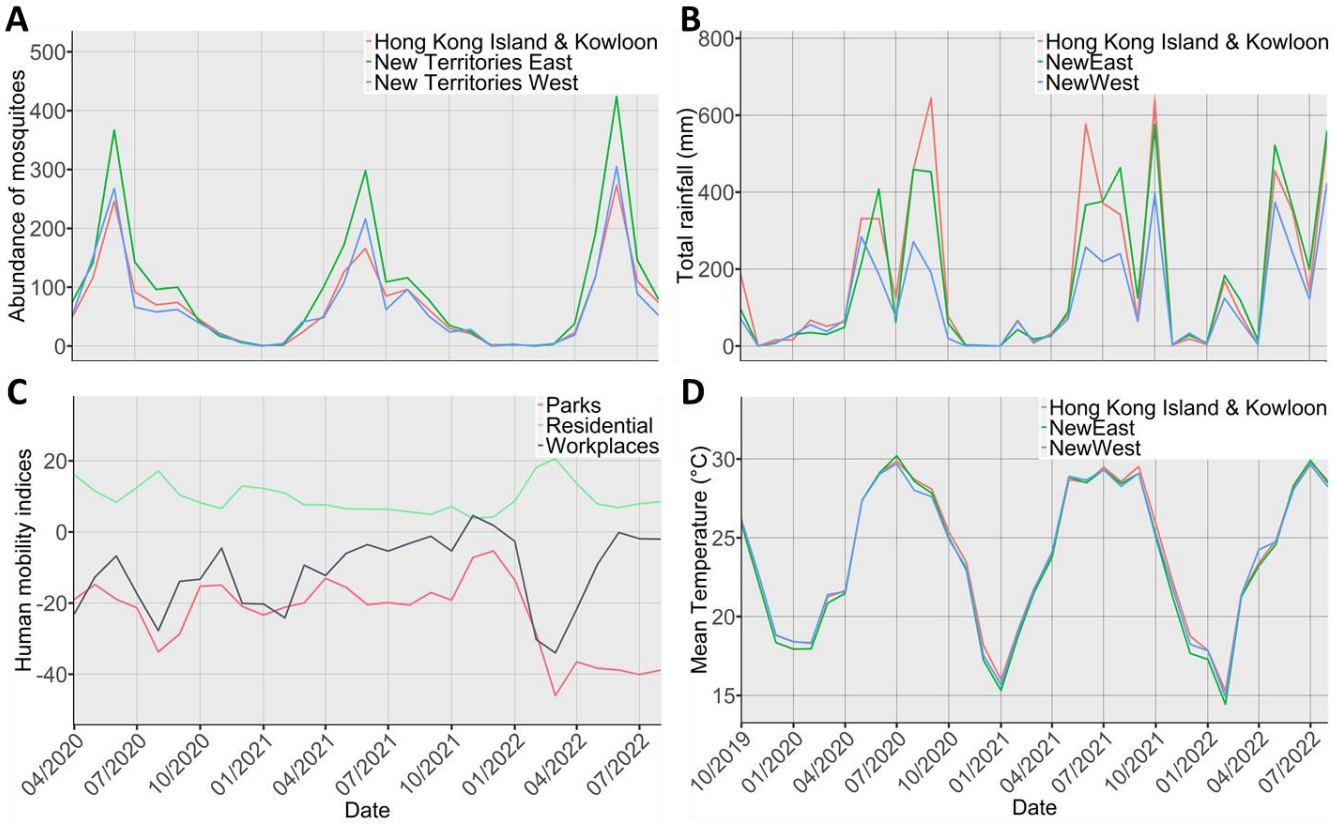


Fig. The monthly change in mosquito abundance and its predictors in Hong Kong. (A) Mosquito abundance; (B) Total rainfall; (C) Human mobility; (D) Mean temperature.

Experimental Results: Prediction results and impact of mobility

- The prediction model performs the best when incorporating human mobility in residential.
- The reduction in the residential index correlated with a rise in mosquito abundance.
- If human mobility (i.e. residential) returned to normal in 2022, the risk of mosquito activity would increase significantly (more than 80%) during the peak.

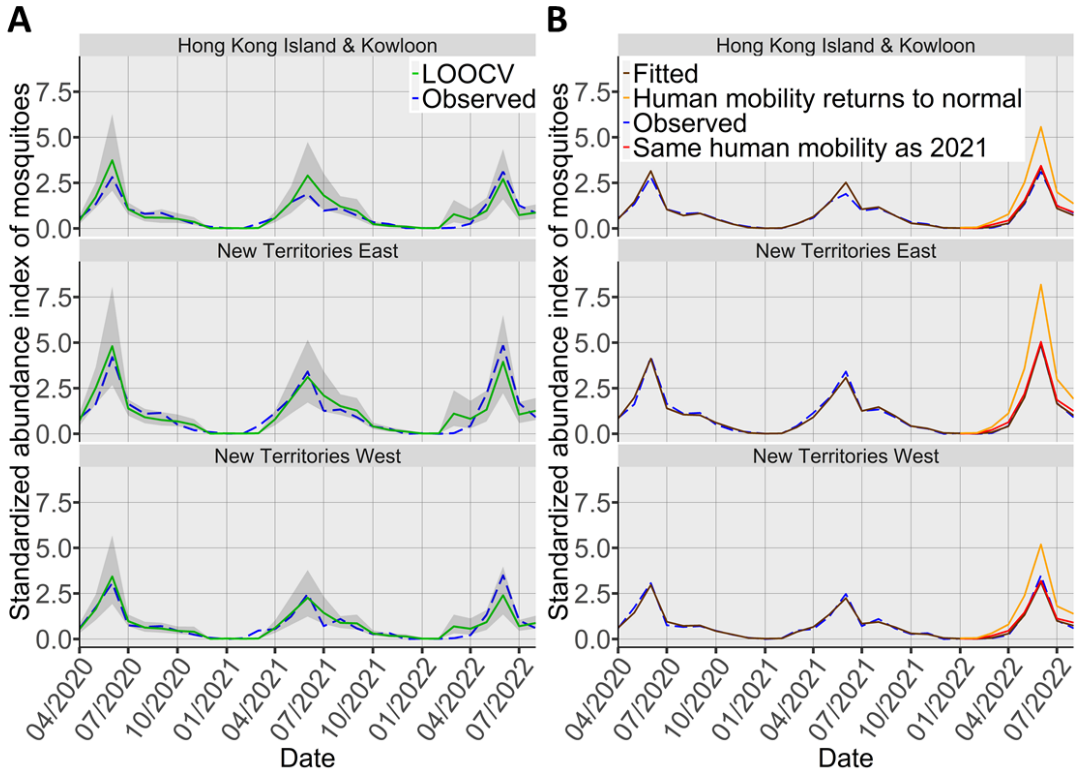


Fig. Comparison of observed and predicted results using the best model for mosquito abundance.

Tab. Comparison of candidate models for mosquito abundance.

Model	Model formula	WAIC	MSE in LOOCV
Model A1	$\beta + \gamma + S + \alpha$	741.12	—
Model A2	$\beta + \gamma + S + f.w(R_t, 6) + \alpha$	727.31	—
Model A	$\beta + \gamma + S + f.w(R_t, 6) + f.w(T_t, 2) + \alpha$	681.60	0.63
Model A-M_p	$\beta + \gamma + S + f.w(R_t, 6) + f.w(T_t, 2) + m_p + \alpha$	677.92	0.65
Model A-M_w	$\beta + \gamma + S + f.w(R_t, 6) + f.w(T_t, 2) + m_w + \alpha$	658.89	0.46
Model A-M_r	$\beta + \gamma + S + f.w(R_t, 6) + f.w(T_t, 2) + m_r + \alpha$	**651.87**	**0.37**
Model A-M_r1	$\beta + \gamma + S + f.w(R_t, 5) + f.w(T_t, 2) + m_r + \alpha$	666.57	—
Model A-M_r2	$\beta + \gamma + S + f.w(R_t, 4) + f.w(T_t, 2) + m_r + \alpha$	668.37	—
Model A-M_r3	$\beta + \gamma + S + f.w(R_t, 3) + f.w(T_t, 2) + m_r + \alpha$	664.26	—
Model A-M_r4	$\beta + \gamma + S + f.w(R_t, 2) + f.w(T_t, 2) + m_r + \alpha$	671.56	—
Model A-M_r5	$\beta + \gamma + S + f.w(R_t, 1) + f.w(T_t, 2) + m_r + \alpha$	669.79	—
Model A-M_r6	$\beta + \gamma + S + f.w(R_t, 6) + f.w(T_t, 1) + m_r + \alpha$	653.22	—

Tab. Mobility coefficient in different prediction models.

Model	Variable	Coefficient (95% CI)
Model A-M_p	m_p	0.0122 (-0.0007, 0.0253)
Model A-M_r	m_r	-0.0753 (-0.1182, -0.0332)
Model A-M_w	m_w	0.0305 (0.0119, 0.0496)

Experimental Results: Effects of weather

Total rainfall:

Heavy rainfall conditions (>500 mm) within 3 months were associated with a higher risk of mosquito activity. On the other hand, low rainfall (<50 mm) was associated with a higher risk with a longer lag of about 4.5 months.

Mean temperature:

Temperature was positively associated with mosquito activity until exceeding a threshold.

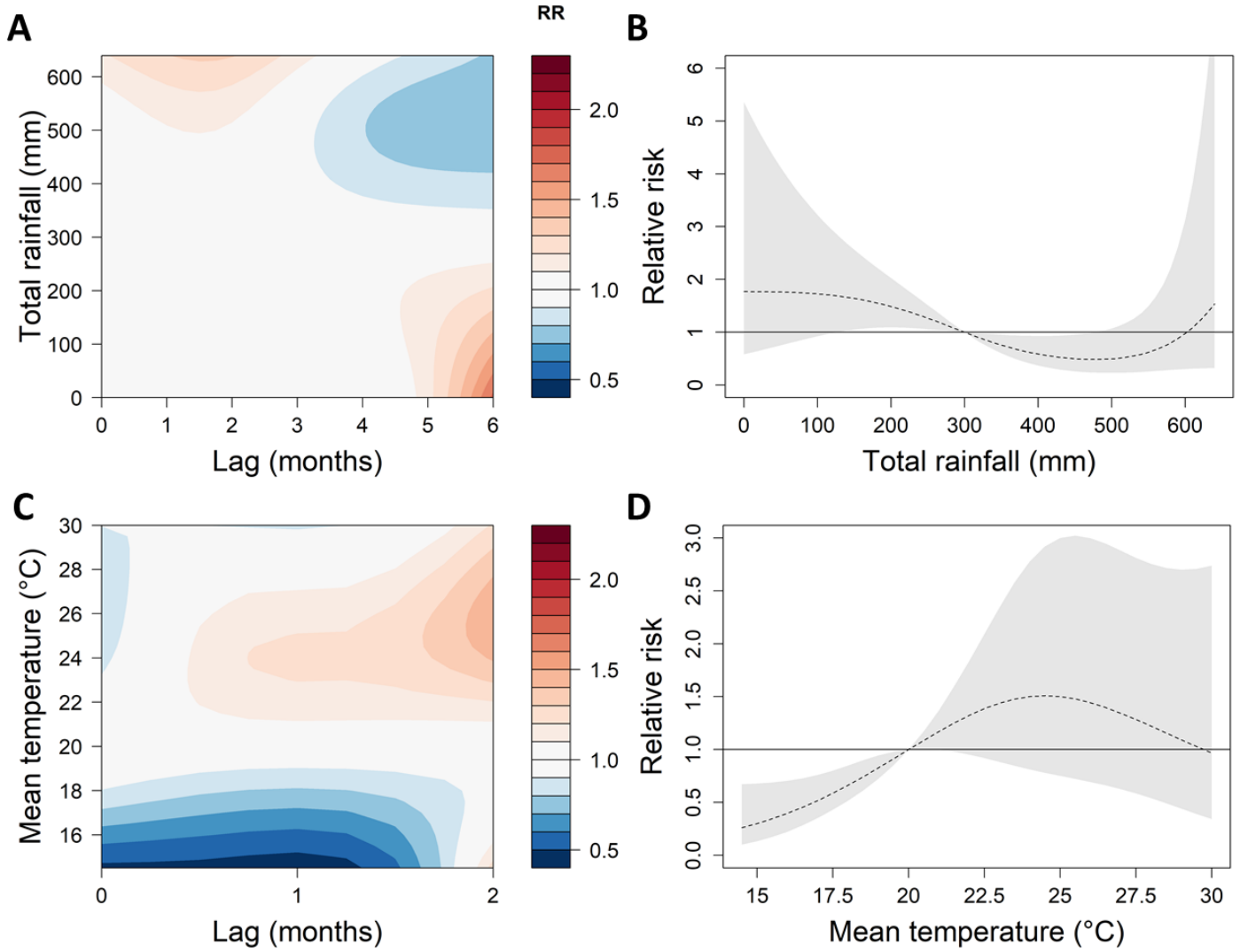


Fig. Effects of total rainfall and mean temperature on mosquito extensiveness using the best model.

Conclusion

- Human mobility is a critical factor in mosquito prediction, which helps to improve the performance of abundance and extensiveness prediction.
- Heavy rainfall and low rainfall both have a positive association with mosquito activity in long lag (>3 months later). Temperature was positively associated with mosquito activity until exceeding a threshold.
- Social distancing measures may be an important intervention to change mosquito abundance and extensiveness.

Thank you!