Supplementary material

Miguel Guzmán-Merino[a], Christian Durán[a], Maria-Cristina Marinescu[b], Concepción Delgado-Sanz[c,d], Diana Gomez-Barroso[c,d], Jesus Carretero[a] and David E. Singh[a,**]

[a] Universidad Carlos III de Madrid, Leganes, Spain
[b] Barcelona Supercomputing Center, Barcelona, Spain
[c] CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[d] National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain

ABSTRACT

As long as critical levels of vaccination have not been reached to ensure heard immunity, and new SARS-CoV-2 strains are developing, the only realistic way to reduce the infection speed in a population is to track the infected individuals before they pass on the virus. Testing the population via sampling has shown good results in slowing the epidemic spread. Sampling can be implemented at different times during the epidemic and may be done either per individual or for combined groups of people at a time. The work we present here makes two main contributions. We first extend and refine our scalable agent-based COVID-19 simulator to incorporate an improved socio-demographic model which considers professions, as well as a more realistic population mixing model based on contact matrices per country. These extensions are necessary to develop and test various sampling strategies in a scenario including the 62 largest cities in Spain; this is our second contribution. As part of the evaluation, we also analyze the impact of different parameters, such as testing frequency, quarantine time, percentage of quarantine breakers, or group testing, on sampling efficacy. Our results show that the most effective strategies are pooling, rapid antigen test campaigns, and requiring negative testing for access to public areas. The effectiveness of all these strategies can be greatly increased by reducing the number of contacts for infected individual.

1. Social and epidemic models

The following tables show the different parameters used to configure the social model used by EpiGraph in our experiments. It is important to highlight that these parameters are related to the demographic and social conditions of each of the considered regions of Spain. In order to synthesize our results, we show the input parameters used for the province of Madrid. A full detailed data of all the experiment will be available in the public data repository used by EpiGraph and published upon the paper acceptation. The data was collected from the Spanish National Statistics Institute (INE) [1]. The population pyramid (not shown in tables) was also collected from the INE for each Spanish province.

Table 1 shows the percentage distribution of each collective and the sizes of the groups considered for each collective. In Table 2 the work collectives are broken down by profession and include the industry, construction, catering services, security, education, health, elderly care, and transportation. Note that some of the professions have specific contact patterns, which are considered in the social model. More specifically, education and elderly care include static contacts with students and elderly people at nursing home, respectively. For catering, security, and health we consider dynamic contacts. Health professionals are also divided into front-line and non-front-line workers. Each one of these two sub-collectives have different types of dynamic contacts. In Table 3 the elderly collective is broken down by sub-collectives: elderly people at home, in day-care centres, and in nursing homes. Table 4 illustrates the family size distribution used in our simulation. Note that this distribution is also different for each province.

Table 5 shows the list of parameters used to model the individual (i.e. agent). We distinguish between static parameters - with constant values- and dynamic parameters - which may change during the simulation. The table also indicates whether the parameter is used during the simulation. Tables 6 and 7 show the R_0 values and transition probabilities for each compartment state considered in the Epidemic model.

** Corresponding author
ORCID(s):

Table 1
Social group distribution for the cities of Madrid province.

School groups					
MinAge	0	MaxAge	19	Percentage	0.1757 %
MinSize	40	MaxSize	200	Percentage males	0.5108 %
Work groups					
MinAge	20	MaxAge	64	Percentage	0.5179 %
MinSize	20	MaxSize	1000	Percentage males	0.4770 %
Stay-at-home, informal meetups groups					
MinAge	20	MaxAge	64	Percentage	0.1194 %
MinSize	1	MaxSize	10	Percentage males	0.4770 %
Elder, informal meetups groups					
MinAge	65	MaxAge	100	Percentage	0.1870 %
MinSize	25	MaxSize	50	Percentage males	0.3905 %

Table 2
Work collective breakdown in professions. Edu. and Elderly-CG stands for education professionals and elderly caregivers, respectively. The percentages are the fraction of each profession among the worker collective. $Size_{min}$ and $Size_{max}$ denote the minimum and maximum sizes of each specific collective. A normal distribution between these two values has been used for setting each group size.

	Industry	**Construction**	**Catering**	**Services**	**Security**	**Edu.**	**Health**	**Elderly-CG**	**Transport**
	30.80%	6.50%	8.80%	24.00%	7.40%	7.50%	6.40%	3.30%	5.30%
$Size_{min}$	1	1	1	1	10	6	10	5	1
$Size_{max}$	30	20	12	8	50	30	30	25	8

Table 3
Elderly collective breakdown in classes. Elderly at home represents the elderly people that live at home and participate in day centres (in our simulations, according to the existing conditions in Spain, day centres were closed during the simulation period so this collective was merged with the elderly-at-home collective). The percentages are the fraction of each class among this collective. $Size_{min}$ and $Size_{max}$ denote the minimum and maximum sizes of each specific collective. A normal distribution between these two values has been used for setting each group size.

	Elderly at home	**Elderly at day-care centre**	**Elderly at nursing home**
	50.6%	46.3%	3.1%
$Size_{min}$	4	10	10
$Size_{max}$	10	30	40

Table 4
Family size distribution for the cities of the Madrid metropolitan area.

Number of members in a family					
1 member	25.50 %	2 members	30.40 %	3 members	20.90 %
4 members	17.70 %	5 members	5.50 %		

Table 5
List of parameters used to model the agent. The column labelled $Type$ indicates whether the parameter is static or dynamic, i.e. it has a constant value during the simulation or its value is may change. All these parameters are implemented but only the used ones determine the infection outcome.

Parameter	**Type**	**Comments**	Used
Age	Static	Individual age	Yes
Gender	Static	Male or female	No
Etnic group	Static	White, black, latino, asian, american indian, others	No
Occupation	Static	Student, worker, elderly people or unemployed	Yes
Occupation group	Static	Profession. See Table 2	Yes
Work on Saturday	Static	If true, the individual works on Saturdays	Yes
Health condition	Dynamic	Factors than can increase the risk of severity synonym	No
Mask use	Dynamic	Mask use, type of mask	Yes
Quarantined	Dynamic	Isolation	Yes
Vaccination type	Dynamic	Vaccine type: Pfizer-Biontech, Moderna, Astra-Zeneca or Janssen.	Yes
Vaccination t1	Dynamic	Vaccination time of the first dose	Yes
Vaccination t2	Dynamic	Vaccination time of the second dose	Yes
Infection stage	Dynamic	If infectious, the infection stage related to the individual. See Table 7	Yes
COVID-19 variant	Dynamic	COVID-19 variant: Wuhan, British, E484K or Delta	Yes
Infection t1	Dynamic	Infection start time	Yes
Infection t2	Dynamic	Infection end time	Yes
Sick time	Dynamic	Time that the individual was on bed because of the illness	Yes
Seroprevalence	Dynamic	Prevalence to SARS-COV-2	Yes
Sequels	Dynamic	Infection sequels	No
Test type	Dynamic	Testing method used	Yes
Test time	Dynamic	Testing time	Yes
Extra daily tests	Static	Extra PCR tests in the strategies	Yes
Quarantine breakers	Static	Percentage of individuals that break quarantine time	Yes
Test window	Static	Days for testing the same individual consecutively	Yes

Table 6
R_0 Values and transition probabilities for each compartment state. In this work we have not considered the use of antivirals, thus I_V^S state is not reached and the associated R_0^{ISV} value is not applicable. E^S and A states do not have a related transition probability because there is only a destination state. P^{A_T} represents the transition to asymptomatic for vaccinated individuals. This probability is vaccination-dependant.

Compartment state	R_0 values		Probability	
E^P	R_0^{EP}	0	P^A	25%
E^S	R_O^{ES}	1.42		100%
A	R_O^A	1.42		100%
I^P	R_O^{IP}	4.5	P^{IS}	100%
I^S	R_O^{IS}	3.38	P^H	Table 7
I_V^S	R_0^{ISV}	N/A		100%
H	R_0^H	0.34	P^D	Table 7
E_T^P	R_0^{EP}	0	P^{A_T}	25%
E_T^S	R_O^{ES}	0 or 1.42		100%
A_T	R_O^A	1.1 or 1.42		100%
I_T^P	R_O^{IP}	0 or 4.5	P^{IS}	100%
I_T^S	R_O^{IS}	0 or 3.38	P^H	Table 7
H_T	R_0^H	0 or 0.34	P^D	Table 7

Table 7
Values of P^H and P^D are based on age. P^H is the probability an infected person has of becoming hospitalized and P^D is the probability a hospitalized person (a fraction of the total infected) has of dying.

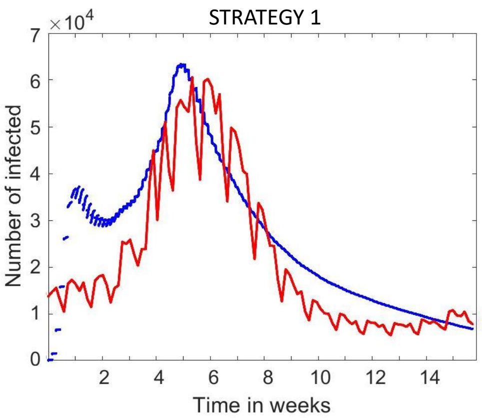
	Age interval								
	< 10	10-19	20-29	30-39	40-49	50-59	60-69	70-79	≥ 80
P^H	0.4%	0.4%	3.4%	9.0%	19.6%	31.4%	40.8%	49.8%	45.2%
P^D	0.0%	0.4%	0.8%	0.8%	1.2%	2.0%	4.7%	12.2%	30.0%

Table 8
R_0 values for the vaccination-related states in each of the considered transmission scenarios.

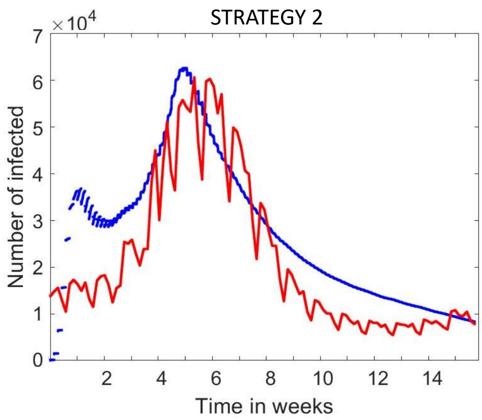
State	A_T	E_T^S	I_T^P	I_T^S	H_T
Non contagious	0	0	0	0	0
partially contagious	0	1.42	4.5	3.38	0.34
fully contagious	1.42	1.42	4.5	3.38	0.34

2. Sampling strategy time plots

Figures 1, 2 and 3 show the time plot of the different strategies. These figures represent the aggregated number of existing infected individuals at a certain time. The simulated and real values are represented in blue and red colors, respectively.



(a)



(b)

Figure 1: Strategies 1 and 2 time plot. The simulated and real values are represented in blue and red colors, respectively. (a) Baseline strategy. (b) Random testing strategy.

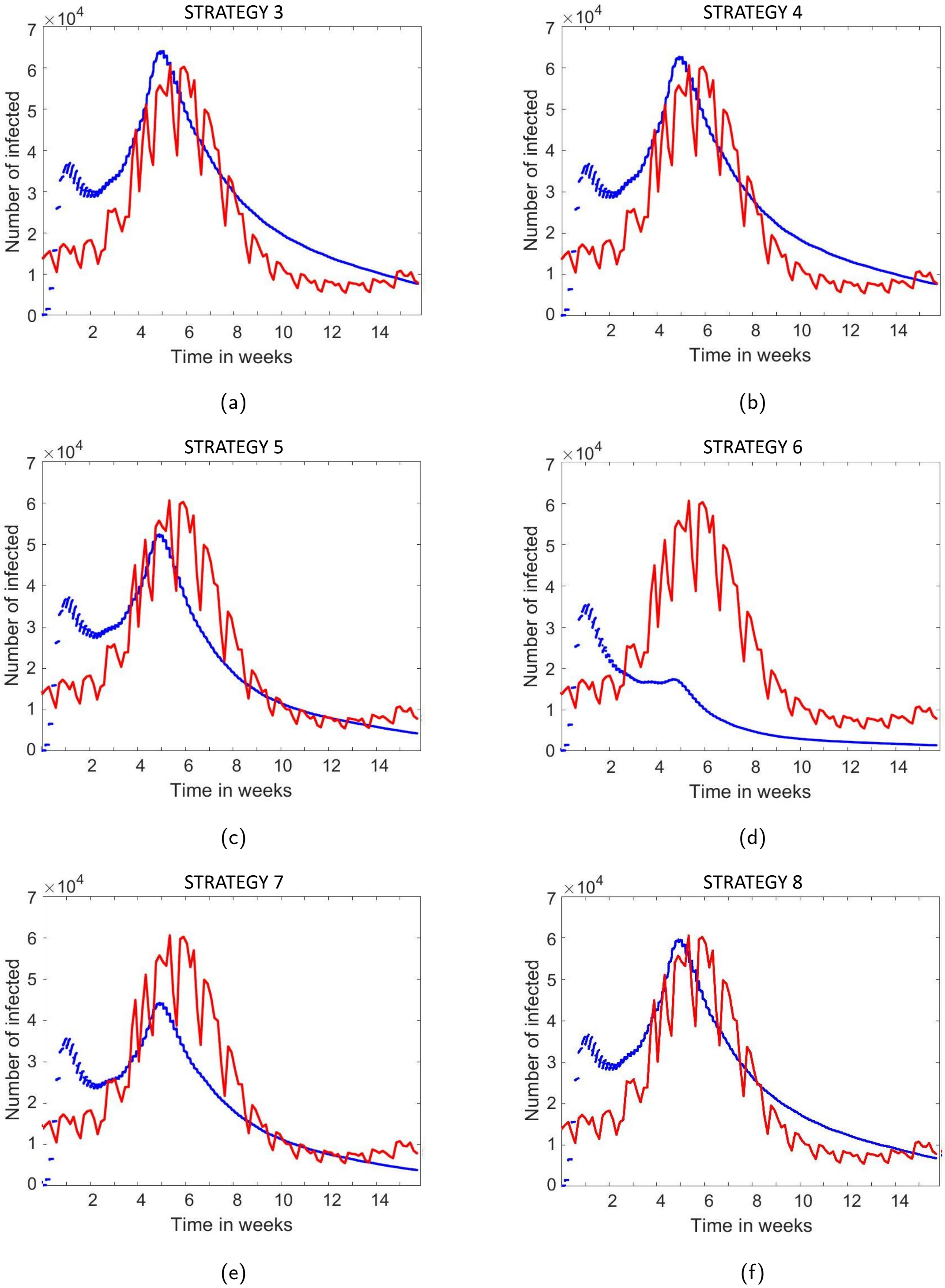


Figure 2: Strategies 3-8 time plot. The simulated and real values are represented in blue and red colors, respectively. (a) Health, social-health, and defense workers testing strategy. (b) Catering workers testing strategy. The Y axis represents that aggregated number of existing infected individuals at a certain time. (c) Infected family contacts quarantined strategy. (d) All contacts of positive individuals are quarantined. (e) Pooling testing method. (f) Pooling testing method in catering workers. The Y axis represents that aggregated number of existing infected individuals at a certain time. The simulated and real values are represented in blue and red colors, respectively.

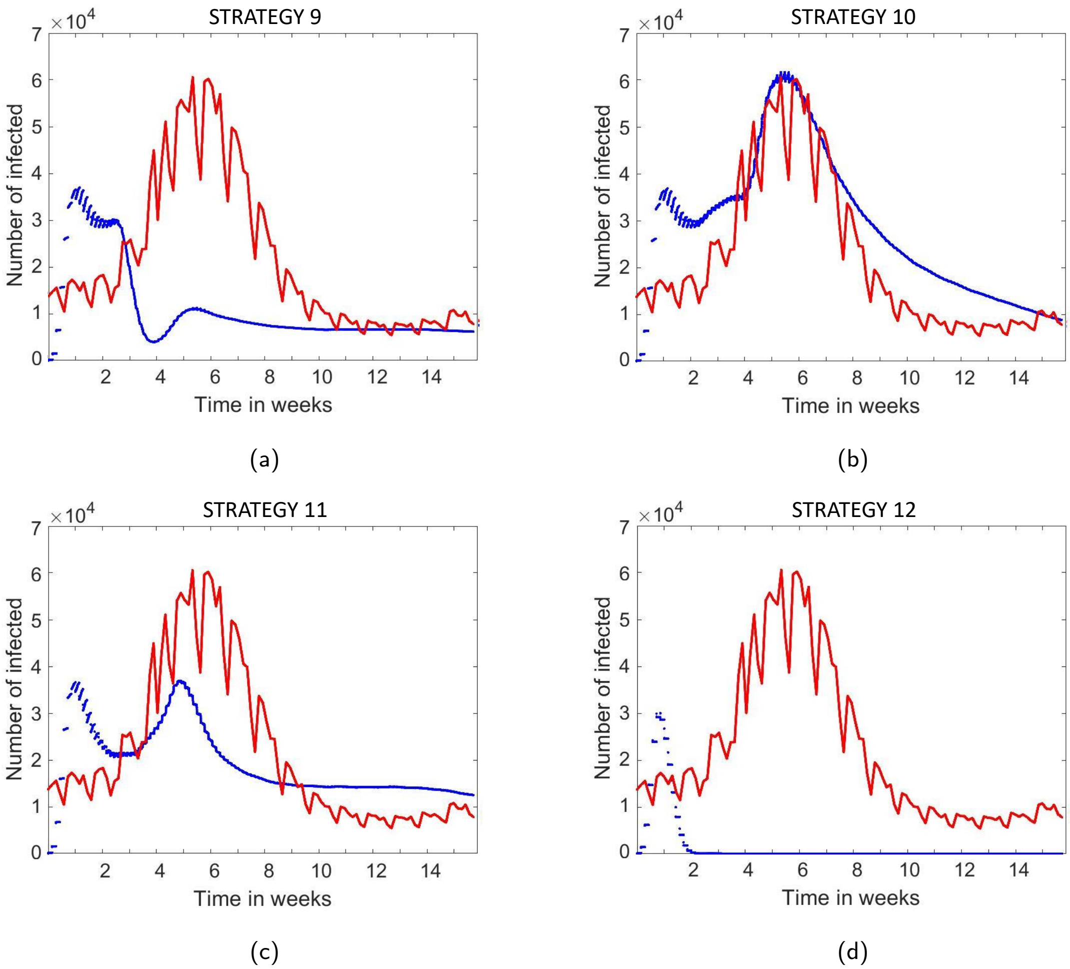


Figure 3: Strategies 9-12 time plot. The simulated and real values are represented in blue and red colors, respectively. (a) Testing campaign. (b) Testing campaign in catering workers. (c) Testing for leisure. (d) Ideal strategy. The Y axis represents that aggregated number of existing infected individuals at a certain time. The simulated and real values are represented in blue and red colors, respectively.

3. Sensitivity analysis

This section tests the robustness of the results of EpiGraph in the presence of uncertainty in the input data. In this study we analyse how the uncertainty in the input parameters used by EpiGraph may affect the results of the simulation.

3.1. Monotonicity analysis

In this section the we present the monotonicity analysis of the main model variables. We have considered the set of parameters that we think are the most relevant in the simulation process. Most of these parameters are depicted in Section 1 of the Supplementary Material. The considered parameters are: R_0^{ES}, R_0^{A}, R_0^{IP}, R_0^{IS}, R_0^{H}, ρ^A, latent secondary period, infected primary period, infection secondary period, hospitalization period, immunity workers percentage, immunity students percentage, immunity unemployed percentage, immunity elderly percentage, initial infected individuals, latent primary period, PCR tests per day, percentage of individuals who break quarantine , and days of quarantine. Figures 4, 5 and 6 show the results of this analysis. Based on these results, the variables chosen for carrying out the sensitivity analysis are: R_0^{ES}, R_0^{A}, R_0^{IP}, R_0^{IS}, ρ^A, immunity workers percentage, immunity students percentage, immunity unemployed percentage, immunity elderly percentage, initial infected individual percentage, percentage of individuals who break quarantine, and number of days of quarantine.

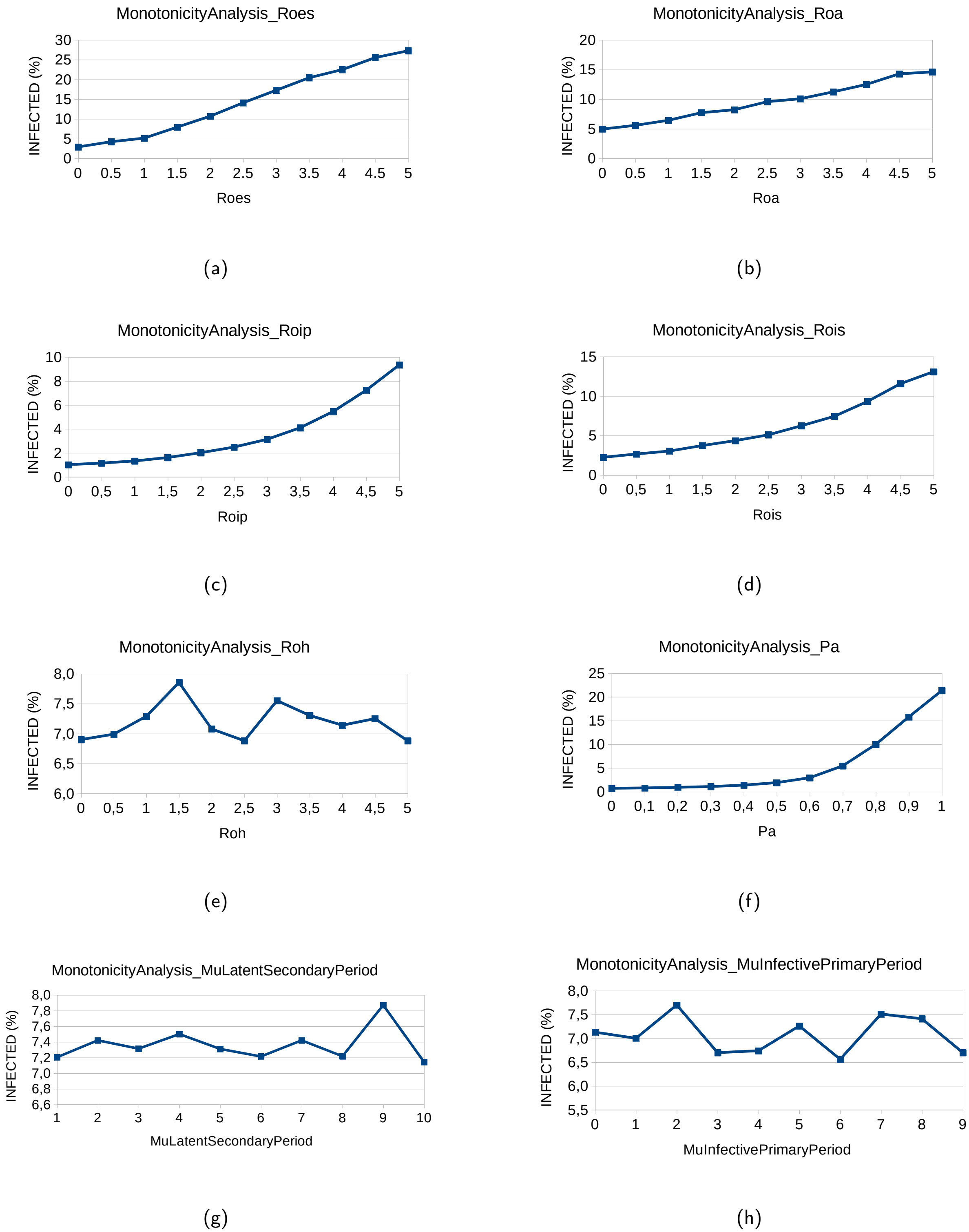
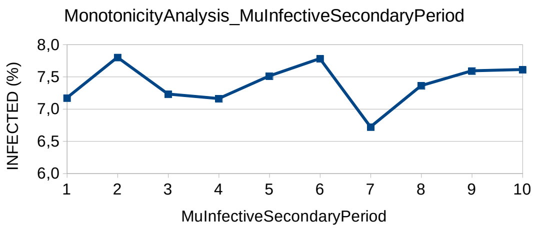
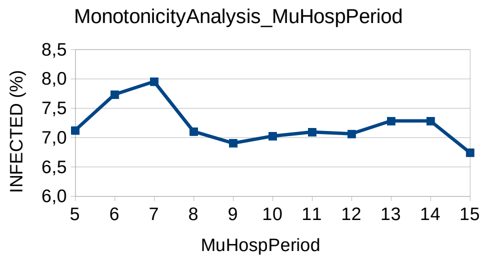


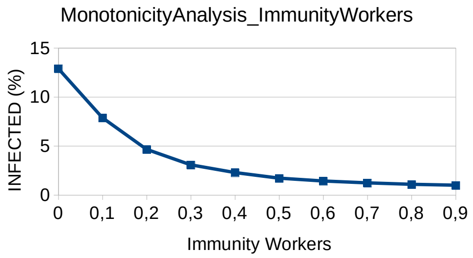
Figure 4: Monotonicity analysis for parameters: (a) R_0^{ES}, (b) R_0^{A}, (c) R_0^{IP}, (d) R_0^{IS}, (e) R_0^{H}, (f) ρ^A, (g) latent secondary period, and (h) infected primary period.



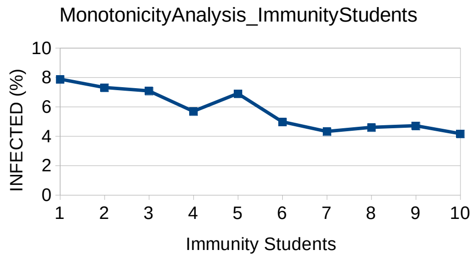
(a)



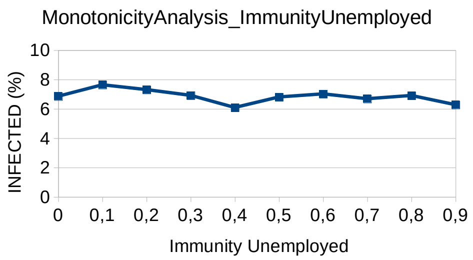
(b)



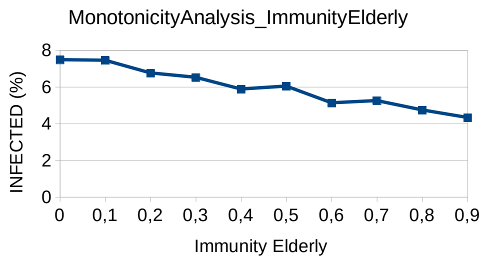
(c)



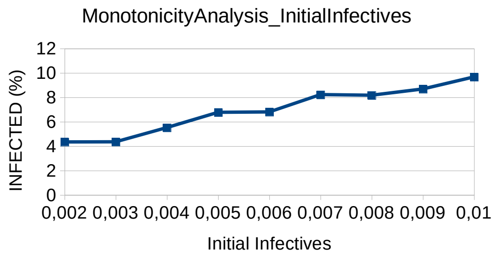
(d)



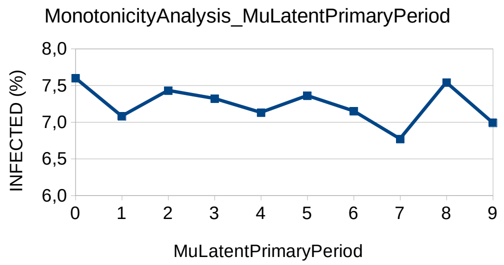
(e)



(f)

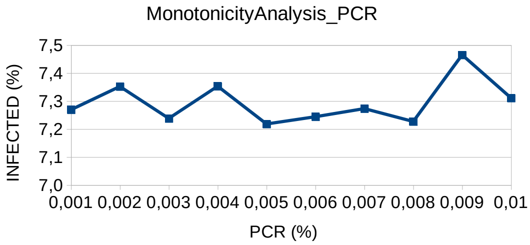


(g)

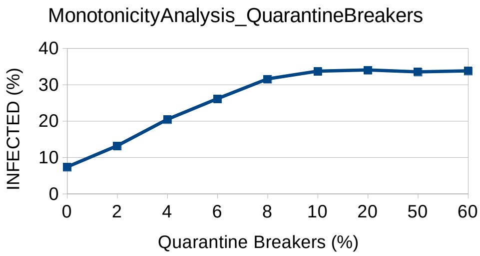


(h)

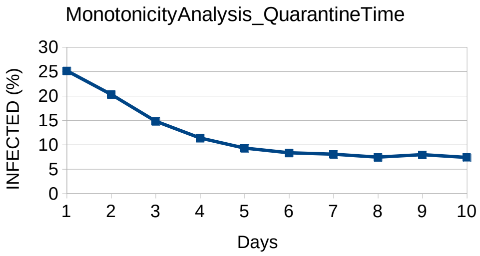
Figure 5: Monotonicity analysis for parameters: (a) infectedt secondary period, (b) hospitalization period, (c) immunity workers percentage, (d) immunity students percentage, (e) immunity unemployed percentage, (f) immunity elderly percentage, (g) Initial infectives, and (h) LatentPrimaryPeriod.



(a)



(b)



(d)

Figure 6: Monotonicity analysis for parameters: (a) PCR tests based on the simulated population, (b) Quarantine breakers, and (c) Quarantine time.

3.2. LHS-PRCC analysis

In order to analyse the influence of the variables on the final result, a Partial Rank Correlation Coefficient analysis is performed. Variables which do not present a monotone relation with the output variable - based on the monotonicity analysis - were excluded from the PRCC analysis. Figure 7 shows the results of this analysis.

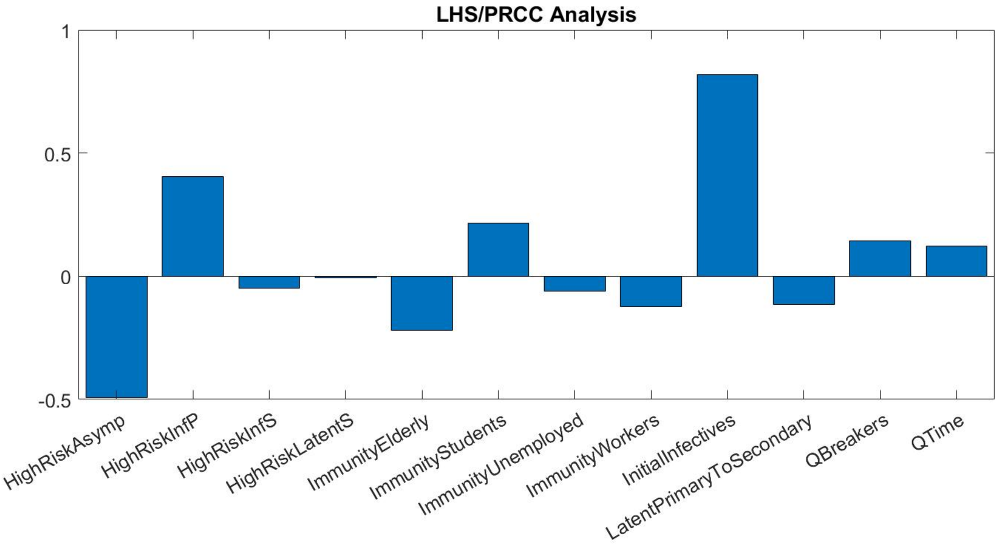


Figure 7: Partial rank correlation coefficient analysis for variables: HighRiskAsymp (R_0^A), HighRiskInfP (R_0^{IP}), HighRiskInfS (R_0^{IS}), HighRiskLatentS (R_0^{ES}), Immunity elderly percentage, Immunity students percentage, Immunity unemployed percentage, Immunity workers percentage, Initial percentage of infected individuals, LatentPrimaryToSecondary (ρ^A), QuarantineBreakers (QBreakers) and QuarantineTime (QTime).

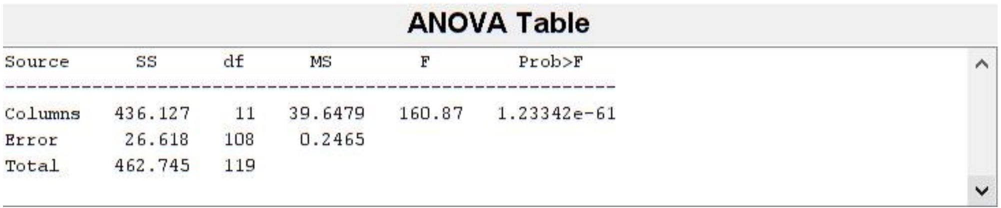
4. Tukey analysis

This analysis is performed in order to identify the main differences among sampling strategies results. In a first stage of this analysis, each of the strategies is simulated 10 times and the final percentage of infected population is obtained using the average values. These values are shown in Table 9.

Table 9
Results of 10 independent simulations for each sampling strategy. The last row of the table shows the average results for each strategy.

Strategy	1	2	3	4	5	6	7	8	9	10	11	12
Test #1	6,09	6,05	6,56	6,55	4,92	2,25	4,48	6,29	2,27	6,68	4,10	0,39
Test #2	6,09	5,63	6,08	6,28	4,62	2,30	3,98	5,74	2,67	6,60	5,17	0,39
Test #3	6,36	6,26	6,07	6,01	4,78	2,25	4,49	5,37	2,56	6,43	4,35	0,39
Test #4	6,18	6,54	6,61	6,22	5,14	2,30	4,48	5,73	2,56	6,58	5,01	0,39
Test #5	6,14	6,53	6,03	6,23	4,90	2,23	3,91	5,98	2,47	6,23	4,56	0,39
Test #6	6,32	6,17	5,87	5,57	5,63	2,29	4,37	6,04	2,56	6,50	4,02	0,39
Test #7	5,78	6,08	6,49	6,34	4,71	2,28	4,59	6,15	2,39	7,03	4,98	0,39
Test #8	5,97	6,12	6,74	6,58	4,44	2,16	4,05	5,77	2,98	6,91	4,63	0,39
Test #9	5,64	6,26	6,31	5,67	5,05	2,18	4,39	5,93	2,29	6,60	4,60	0,39
Test #10	6,31	6,84	6,25	5,97	4,78	2,12	4,14	5,62	2,59	6,63	4,49	0,39
Mean	6,09	6,25	6,30	6,14	4,90	2,24	4,29	5,86	2,58	6,62	4,60	0,39

In order to find the mean square error, an Anova analysis is performed using the Matlab *anova*1 function shown in Figure 8.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	436.127	11	39.6479	160.87	1.23342e-61
Error	26.618	108	0.2465		
Total	462.745	119			

Figure 8: Matlab Anova analysis result of 10 independent simulations for each sampling strategy.

The Tukey criterion is defined by the following equation:

$$[h!]T = q_\alpha(c, n-c)\sqrt{\frac{MSE}{n_i}} = 4.353\sqrt{\frac{0.2465}{120}} = 0.6834 \quad (1)$$

where

- $q_\alpha(c, n-c)$ is the studentized range distribution
- c is the number of treatments
- n is the number of total samples
- MSE is the mean squared error (Anova)
- n_i is the sample size of the treatments group

The Tukey analysis establishes that when the absolute mean differences between groups is less than the Tukey criterion (T) there is not evidence to say that those groups are different. On the other hand, when the absolute mean differences between groups is more than T, it can be said that the groups are different. The format of the results provided by the Tukey analysis is depicted in Table 10, This table illustrates the absolute difference between group means.

Table 11 shows the Tukey analysis results for the considered sampling strategies. The results highlighted in green are larger than the critical value of the Tukey criterion T=0.6834. This means that there is statistical significance between the strategy results.

Table 10
Table format for the Tukey analysis results.

Strategy	1	2	...
1	$\lvert x1 - x1 \rvert$	$\lvert x1 - x2 \rvert$	
2	$\lvert x2 - x1 \rvert$	$\lvert x2 - x2 \rvert$	
...			...

Table 11
Tukey analysis results for the considered sampling strategies.

Strategy	1	2	3	4	5	6	7	8	9	10	11	12
1	0,00	0,16	0,21	0,05	1,19	3,85	1,80	0,23	3,50	0,53	1,49	5,70
2	0,16	0,00	0,05	0,11	1,35	4,01	1,96	0,39	3,66	0,37	1,65	5,86
3	0,21	0,05	0,00	0,16	1,40	4,07	2,01	0,44	3,72	0,32	1,70	5,91
4	0,05	0,11	0,16	0,00	1,25	3,91	1,85	0,28	3,56	0,48	1,54	5,75
5	1,19	1,35	1,40	1,25	0,00	2,66	0,61	0,97	2,31	1,72	0,29	4,51
6	3,85	4,01	4,07	3,91	2,66	0,00	2,05	3,63	0,35	4,38	2,37	1,85
7	1,80	1,96	2,01	1,85	0,61	2,05	0,00	1,57	1,70	2,33	0,31	3,90
8	0,23	0,39	0,44	0,28	0,97	3,63	1,57	0,00	3,28	0,76	1,26	5,47
9	3,50	3,66	3,72	3,56	2,31	0,35	1,70	3,28	0,00	4,04	2,02	2,19
10	0,53	0,37	0,32	0,48	1,72	4,38	2,33	0,76	4,04	0,00	2,02	6,23
11	1,49	1,65	1,70	1,54	0,29	2,37	0,31	1,26	2,02	2,02	0,00	4,21
12	5,70	5,86	5,91	5,75	4,51	1,85	3,90	5,47	2,19	6,23	4,21	0,00

References

[1] Ministry of Economic Affairs and Digital Transformation (MINECO), http://www.ine.es/ (accessed 3 July 2021). *National Statistics Institute (INE)*, 2021.