ISSN 2316-9281

Scientific Electronic Archives
Issue ID: Vol.19 (4), Jul/Aug 2026, p. 1-5
DOI: http://dx.doi.org/10.36560/19420262214
+ Corresponding author: koscianski@utfpr.edu.br

Agent-based simulation of epidemic incorporating real data

André Koscianski & Márcio Douglas Penteado da Silva

Universidade Tecnológica Federal do Paraná

Abstract. The spread of infectious diseases in large urban populations is strongly influenced by individual-level contact heterogeneity and network structure. Agent-based models combined with network representations can capture this complexity beyond the homogeneous mixing assumptions of classical compartmental models. We developed an agent-based simulator using a Watts-Strogatz small-world network to represent interaction, and variable contact probabilities to reflect patterns observed in data. The model was programmed to generate execution logs summarizing agent histories. The software was tested with actual data from a covid outburst in Curitiba, Brazil, a city of approximately 1.77 million inhabitants, covering a period of 240 days. Contact probabilities were adjusted in eight intervals optimised against actual data using the nomadoptimization tool. Three counterfactual scenarios were evaluated: public transport restrictions, maintained public transport capacity, and delayed school closure. The model achieved a good fit to actual data (RMSE = 9125; R^2 = 0.767) and sensitivity analysis showed that contact probability perturbations during the first 150 days produced the largest effects. The contact probabilities exhibited a temporal pattern consistent with changes in social behaviour and policy interventions. Counterfactual experiments showed strong nonlinear responses. The model demonstrated that the integration of demographic realism, small-world network topology, stochastic simulation, and data-driven calibration produced a computationally feasible framework that reproduces empirical epidemic dynamics of a large urban setting. The simulation outputs (contagion trees, secondary-case distributions, serial intervals) provide epidemiological information not accessible from compartmental approaches, supporting decision making for public health planning and epidemic preparedness.
Keywords: epidemics, computer simulation, agent based models, small-world networks.

Introduction

Mathematical and computational models have proven essential for epidemic decision-making, providing early warning, situational awareness, and scenario testing (Diallo et al. 2021). Such tools help understand complex transmission dynamics, and forecast scenarios in situations with limited data availability (Vespignani et al., 2020; Silva et al., 2020).

Compartmental models such as SEIR offer mathematical tractability, analytical solutions under simplifying assumptions, and computational efficiency that enables rapid exploration of parameter space. This advantages are negatively impacted as they impose homogeneous mixing, requiring all individuals to be equally likely to make contact (Duan et al. 2015). Agent-based models (ABMs) overcome this by capturing heterogeneous contact patterns across households, workplaces, and communities, enabling individual-level interventions such as contact tracing that are impossible in classical frameworks (Ajelli et al. 2010; Willem et al. 2017; Firth et al. 2021).

Network structure is central to epidemic dynamics. Small-world networks with high local clustering combined with occasional long-range links offer a good representation of human social interactions (Watts & Strogatz, 1998; Lloyd et al. 2006). The average degree and rewiring probability are particularly important parameters (Volz et al. 2011). Stochastic simulation further enriches ABMs: multiple runs characterise the full distribution of epidemic trajectories and enable uncertainty quantification impossible with deterministic models (Ahmed et al. 2013; Fadikar et al. 2024). Finally, detailed transmission logs provide outputs not available from aggregate compartmental

approaches, allowing to analyze contagion trees and secondary-case distributions (Firth et al. 2021).

Calibrating ABMs to real data is crucial for their reliability, but this is a challenging task (Marion et al. 2022). ABMs present obstacles such as high dimension of the search space, discontinuous character of the output, and the duration of execution (Audet et al. 2022; Adiga et al. 2025). This motivates the use of specially tailored derivative-free optimization algorithms (Audet et al. 2022).

This study present an agent-based Monte Carlo simulator, calibrated and tested with real data of the covid epidemic in Curitiba, Brazil, a city with 1.77 million inhabitants. The study demonstrates the integration of network structure, stochastic dynamics, and real-data calibration at city scale, allowing to evaluate counterfactual intervention scenarios. This article contributes with methodological insights into calibrating stochastic network models with deterministic optimization.

Methods

The model represents 1770000 agents whose demographic composition (age, sex) follows the 2010 census (IPARDES 2026). Agents were connected through a Watts-Strogatz small-world network with average degree 30 and 0.5% rewiring probability, producing approximately 50 million edges. The simulation kept records of individual data such as sex, age, date of contamination, and also propagation chains (Großmann et al. 2021; Bostanci & Conrad, 2025).

Disease progression followed timelines obtained from the literature (Ferretti et al. 2020). Each day, every infectious agent tests all network neighbors for contact; if contact is established, age- and sex-specific contamination rates determine whether transmission occurs (UFPR, 2026). The contact test is governed by a probability parameter (PC) that can differ across eight consecutive 30-day intervals, capturing temporal variation in transmission intensity consistent with changes in R^2 (Moein et al. 2021). The initial infected count was set to 10 after initial verification tests.

Calibration minimized the mean squared error between simulated and actual cumulative case curves using the NOMAD derivative-free optimizer (Audet et al. 2022). Each calibration run used 10 Monte Carlo iterations. Final outputs used 50 Monte Carlo runs. A sensitivity analysis (Broeke et al., 2016) perturbed each of the eight PC values by ±1%, ±2%, and ±3% independently, to identify critical intervention windows.

Counterfactual scenarios

One of the main advantages of agent based models is the extent to which characteristics of the simulation can be modified in order to analyze its behavior (Tracy et al., 2018). In the case of the epidemic used in this study, following Brazil's Federal Emergency declaration (4 Feb 2020), Curitiba closed schools (23 Mar) and commercial activities (26 Mar, Decree 470/2020). Transport capacity was reduced starting at May; ridership fell from 1.3 million in 2019 to 0.7in 2020. In this context, three scenarios were tested: (i) public transports slightly more active; (ii) earlier restrictions on public transportations; (iii) keeping schools open for one additional month. Each scenario used 50 Monte Carlo runs; pair wise comparisons with the baseline used Welch's t-test and Cohen's d.

Results

Calibration and sensitivity

The calibration of the simulator yielded the following values for the probability of contact: 0.03 0.025 0.04 0.018 0.03 0.011 0.02 0.01, a substantial temporal variation across the eight 30-day periods; it peaked in days 60-90 (onset of exponential growth), fell sharply in days 90-120 (intensified behavioral change), then declined steadily. The best single run achieved RMSE = 2333, R^2 = 0.9, and 55,813 total cases. All PC values were below 4%, reflecting compensation for the dense network structure where even low probabilities generate sufficient transmission.

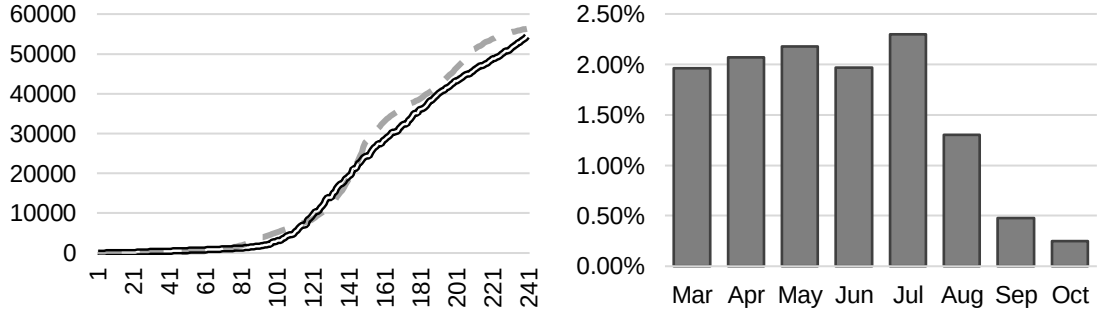

Figure 1. Simulation results. The left plot shows the actual data as a double black line, and the simulation as a single grey line. The right plot shows the sensitivity analysis for the eight Probability Contagion (PC) values.

Sensitivity was approximately linear at all perturbation levels. Absolute sensitivity for periods 1-5 (days 0-150) averaged 1,159 additional cases per 1% PC change (range 1,087-1,271). Sensitivity fell sharply for periods 6-8, reaching 719, 266, and 140 cases respectively. This confirms that contact-reducing interventions are most impactful during the exponential growth phase; measures after day 150 have substantially reduced efficacy.

Simulated epidemic characteristics

The basic reproduction number was R_0 = 1.765, placing the simulated pathogen in the range estimated for seasonal influenza and early COVID-19 variants. The secondary-case distribution was right-skewed and consistent with a negative binomial pattern: 43% of infected agents caused no secondary infections (noting the simulation ended at day 240); at the tail, one individual infected 10 others. There were 23510 root-to-leaf transmission chains with a maximum length of 35 generations, a mean of 20.8, and a median of 21, corresponding to a generation interval of approximately 9.2 days.

The age profile of new cases formed a bell-shaped distribution centred on the 30-39 age group. Working-age adults between 20 and 59 account for 70% of all transmission events. This agrees with empirical contact-rate data, since individuals in this age range have the highest rate of social and occupational interactions.

Analysis of counterfactual scenarios

With the calibrated simulator, it was possible to evaluate alternative scenarios for the progression of the epidemic. The tests mentioned in the methodology session were configured based on sensitivity analysis, as follows: (i) public transports more active: PC was increased by 0.5% from May until the end of the simulation window; (ii) earlier restrictions on transports: PC decreased by 0.5% during the first two months; (iii) schools open one additional month: PC was increased by 50% in April, but only for individuals with ages between 10 and 17.

Each scenario was simulated with 50 Monte Carlo runs. Pair wise comparisons between each scenario and the baseline were performed using Welch's two-sample t-test, which does not assume equal variances. Test statistic, two-tailed p-value, and Cohen's d effect size were computed; effect sizes were interpreted following conventional thresholds: $|d| < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, >0.8 large. The Table 1 shows a summary of results.

Scenario 1 compressed the epidemic vertically without shifting its timing, demonstrating that a modest persistent reduction in individual contact risk can push the system below the epidemic threshold. The 60.2% case reduction for a 0.5% probability change is a direct signature of supercritical nonlinearity.

The second Scenario produced an asymmetrically larger response to the same perturbation magnitude: by month three, the restored transport capacity acted on a substantially larger infected pool, amplifying cases exponentially and delaying the peak by nearly two months.

Scenario 3 showed that even a temporary, targeted increase in school-age contact had irreversible network effects — the 25.7% excess persisted long after schools were closed, because early seeding events propagated permanently through connected households and workplaces. Across all scenarios, peak timing was stable to within one day while final case counts varied by up to 28% coefficient of variation, indicating that epidemic timing is predictable but epidemic magnitude is not.

Table 1. Summary of statistics for the baseline simulation calibrated with real data and the three counterfactual scenarios.

Test	Final cases	Std. dev.	95% CI	P5–P95	Peak day	Peak	Dif. base-line	Welch- t	p-value	Cohen d
Baseline	56305	15815	[51922–60689]	[28304–80087]	148.6	1205	–	–	–	–
1	22384	8694	[19974–24794]	[9480–38422]	148.5	479	-60%	13.29	<0.001	2.66
2	356955	101919	[328705–385206]	[191534–502358]	208.6	6192	534%	-20.61	<0.001	-4.12
3	70760	21688	[64749–76772]	[39562–100187]	148.8	1524	25%	-3.81	<0.001	-0.76

Taken together, the three counter-factual scenarios reveal several structural properties of the simulated epidemic system.

First, the system exhibits pronounced nonlinear sensitivity to the global contagion probability. This is a direct consequence of the supercritical epidemic regime: small changes in the effective reproduction number translate into large differences in final epidemic size through the exponential dynamics of the early growth phase.

Second, there is a strong asymmetry in the system's response to increases versus decreases in transmissibility. Two factors are involved: the multiplicative nature of epidemic spread and the

delayed onset of the Scenario 2 perturbation, since there were already a significant number of patients.

Third, Scenarios 1 and 3 confirm the epidemiological principle that interventions preventing early transmission are more efficient and tend to exceed the costs of delayed action. The entire 60% reduction achieved in Scenario 1 results from a persistent but modest reduction in individual contact risk, while Scenario 3 shows that even one month of elevated transmission in a subgroup increases the total result. The difference lies in timing: the transmission boost among students occurred during the early exponential phase, when it was maximally likely to seed new transmission chains.

Fourth, the model reveals interesting stochastic properties. The coefficient of variation of final case counts in the baseline was approximately 28%, indicating substantial run-to-run variability; this was amplified in all what-if scenarios. By contrast, the peak day was stable to within one day across all scenarios and all runs. This pattern suggests that the dynamics of the epidemics is mostly governed by the network structure and the time-varying contagion schedule, while stochastic variation in early seeding governs the scale of the trajectories. For policy purposes, the stability of epidemic timing across scenarios means that the timing of healthcare demand peaks is highly predictable, whereas the magnitude of that demand is subject to substantial uncertainty.

Conclusions

The verification of the model confirmed the suitability of the approach, which was to embed real epidemic data into the behavior of a computational simulator. The change in contact probabilities over time does not claim to offer a causal explanation but is anchored in reality as it echoes the effect of several fluctuating parameters present in the phenomenon. The calibrated model successfully reproduced the COVID-19 epidemic trajectory in Curitiba ($R^2 = 0.9$), confirming that agent-based simulation on a Watts-Strogatz network, anchored to census demographics and optimised against real data, provides a tractable framework for large-scale urban epidemics. The temporal pattern of calibrated contact probabilities encoded the aggregate effect of successive non-pharmaceutical interventions without requiring causal attribution to individual measures.

Four structural properties of the simulated epidemic system emerge from the results. First, pronounced nonlinear sensitivity to contact probability: changes of less than 1% translate into order-of-magnitude differences in final case counts near the epidemic threshold. Second, strong asymmetry between upward and downward perturbations, driven by the multiplicative nature of supercritical spread and the timing of the perturbation relative to the infected pool size. Third, the disproportionate value of early intervention: the sensitivity analysis and Scenario 1 jointly demonstrate that the window of high-impact action is front-loaded, narrowing sharply after day 150. Fourth, predictable timing but uncertain scale: the peak day was stable across all scenarios and runs while final case counts were highly variable, a pattern governed by network topology and the contagion schedule rather than stochastic seeding.

Beyond aggregate fit, the detailed transmission outputs — contagion trees, secondary-case distributions, and generation intervals — provide epidemiological information not accessible from compartmental models, enabling validation against empirical transmission parameters and supporting the design of contact-tracing and isolation policies.

Key limitations include the static contact network (real networks evolve dynamically), the single-strain assumption, and calibration against reported cases subject to underreporting. Future work should explore dynamic networks, explicit location-based contact structures, and more efficient calibration methods such as genetic algorithms or Bayesian surrogates. The framework is readily transferable to other pathogens, cities, and intervention configurations.

References

Adiga A, et al. Advancing calibration for stochastic agent-based models in epidemiology with Stein variational inference and Gaussian process surrogates. arXiv:2502.19550; 2025.

Ahmed A, Greensmith J, Aickelin U. Variance in System Dynamics and Agent Based Modelling Using the SIR Model of Infectious Disease. arXiv:1307.2001; 2013.

Ajelli M, Gonçalves B, Balcan D, et al. Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infect Dis. 2010;10:190.

Audet C, Le Digabel S, Montplaisir VR, Tribes C. Algorithm 1027: NOMAD Version 4: Nonlinear Optimization with the MADS Algorithm. ACM Trans Math Softw. 2022;48(3):1-22.

Broeke, G., Van Voorn, G., Ligtenberg, A. Which sensitivity analysis method should I use for my agent-based model?. Journal of Artificial Societies and Social Simulation, 2016;19(1), 5.

Bostanci I, Conrad T. Integrating Agent-Based and Compartmental Models for Infectious Disease Modeling: A Novel Hybrid Approach. JASSS. 2025;28(1):5.

Diallo SY, Wildman WJ, Hodulik G, et al. Computational simulation is a vital resource for navigating the COVID-19 pandemic. Simul Healthc. 2021;16(5):297-310.

Duan W, Fan Z, Zhang P, Guo G, Qiu X. Mathematical and computational approaches to

epidemic modeling: a comprehensive review. Front Comput Sci. 2015;9(5):806-26.

Fadikar A, et al. Towards improved uncertainty quantification of stochastic epidemic models using Sequential Monte Carlo. arXiv:2402.15619; 2024.

Ferretti L, Ledda A, Wymant C, Zhao L, Ledda V, Abeler-Dörner L, et al. The timing of COVID-19 transmission. MedRxiv. 2020;2020-09.

Firth JA, Hellewell J, Klepac P, et al. Using a real-world network to model localized COVID-19 control strategies. Nat Med. 2021;27:1015-1021.

Großmann G, Backenköhler M, Wolf V. Heterogeneity matters: Contact structure and individual variation shape epidemic dynamics. PLoS One. 2021;16(7):e0250050.

IPARDES, Instituto Paranaense de Desenvolvimento Econômico e Social. URL https:// www.ipardes.pr.gov.br/ Pagina/ Tabelas-Censos-Demograficos.

Marion G, Hadley L, Isham V, et al. Challenges for modelling interventions for future pandemics. Epidemics. 2022;38:100546.

Moein S, Nickaeen N, Roointan A, et al. Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. Sci Rep. 2021;11(1):4725.

Silva P. C. L., Batista P. V. C., Lima H. S., COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. Chaos Solitons Fractals. 2020;139:110088.

Tracy M, Cerdá M, Keyes KM. Agent-Based Modeling in Public Health: Current Applications and Future Directions. Annu Rev Public Health. 2018;39(1):77-94.

UFPR, Universidade Federal do Paraná. Dados abertos. URL https://dadosabertos.c3sl.ufpr.br / curitiba / CasosCovid19.

Vespignani A, Tian H, Dye C, et al. Modelling COVID-19. Nat Rev Phys. 2020;2:279-281.

Volz EM, Miller JC, Galvani A, Meyers LA. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol. 2011;7(6):e1002042.

Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998; 393 (6684):440-2.

Willem L, Verelst F, Bilcke J, et al. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC Infect Dis. 2017;17:612.