Supplementary Information: Unveiling the paths of COVID-19 in a large city based on public transportation data

Jorge L. B. Araújo[1*], Erneson A. Oliveira[1,2,3], José S. Andrade Jr.[4], Antonio S. Lima Neto[5,6], Vasco Furtado[2,7]

[1] *Laboratório de Ciência de Dados e Inteligência Artificial, Universidade de Fortaleza, Fortaleza, Ceará, 60811-905, Brasil.*

[2] *Programa de Pós Graduação em Informática Aplicada, Universidade de Fortaleza, Fortaleza, Ceará, 60811-905, Brasil.*

[3] *Mestrado Profissional em Ciências da Cidade, Universidade de Fortaleza, Fortaleza, Ceará, 60811-905, Brasil.*

[4] *Departamento de Física, Universidade Federal do Ceará, Fortaleza, Ceará, 60455-760, Brasil*

[5] *Célula de Vigilância Epidemiológica, Secretaria Municipal da Saúde, Fortaleza, Ceará, 60810-670, Brasil.*

[6] *Centro de Ciências da Saúde, Universidade de Fortaleza, Fortaleza, Ceará, 60811-905, Brasil.*

[7] *Empresa de Tecnologia da Informação do Ceará, Governo do Estado do Ceará, Fortaleza, Ceará, 60130-240, Brasil.*

[*] *Correspondence to: jorgearaujo@unifor.br*

(Dated: January 12, 2023)

I. ANALYSIS OF THE BEST THRESHOLD n^*_{ac}

Figure SI-1 shows the analysis of the best threshold n^*_{ac} for the number of accumulated reported cases n_{ac}. Given a neighborhood i, labeled as the Initial Outbreak Location (IOL), we calculate the Pearson correlation coefficient [1]

$$r_i = \frac{\mathrm{cov}(T_a, D)}{\sigma(T_a)\sigma(D)}, \tag{1}$$

where T_a is the arrival time of the diseases in neighborhood j, D is the shortest path distance from all other neighborhoods j to i, cov is the covariance and σ is the standard deviation. The D is calculated taking into account the average of the weekly Origin-Destination (OD) matrices $M^{(k)}$ from $k = 1$ to $k = 10$, *i.e.*, until the week that the disease reached the last neighborhood for $n^*_{ac} = 20$. Figure SI-1(a) shows the boxplot representation of the r distribution for all possible IOL versus the several values of n^*_{ac}. Therefore, we choose $n^*_{ac} = 6$ as the best threshold value because it is the value that establishes a plateau and maximizes the median of the r distribution. We also show the spatial distribution of the Pearson correlation coefficient r for $n^*_{ac} = 6$ in Fig. SI-1(b). We also note that the most likely neighborhoods to have started the first COVID-19 outbreak in the city of Fortaleza are those from the northern region of the city. Such a result is consistent with observational data.

---

[1] Kutner MH, Nachtsheim CJ, Neter J, Li W. Inferences in regression and correlation analysis. In: Applied Linear Statistical Models. 5th ed. Singapore: McGraw-Hill/Irvin, 2005.

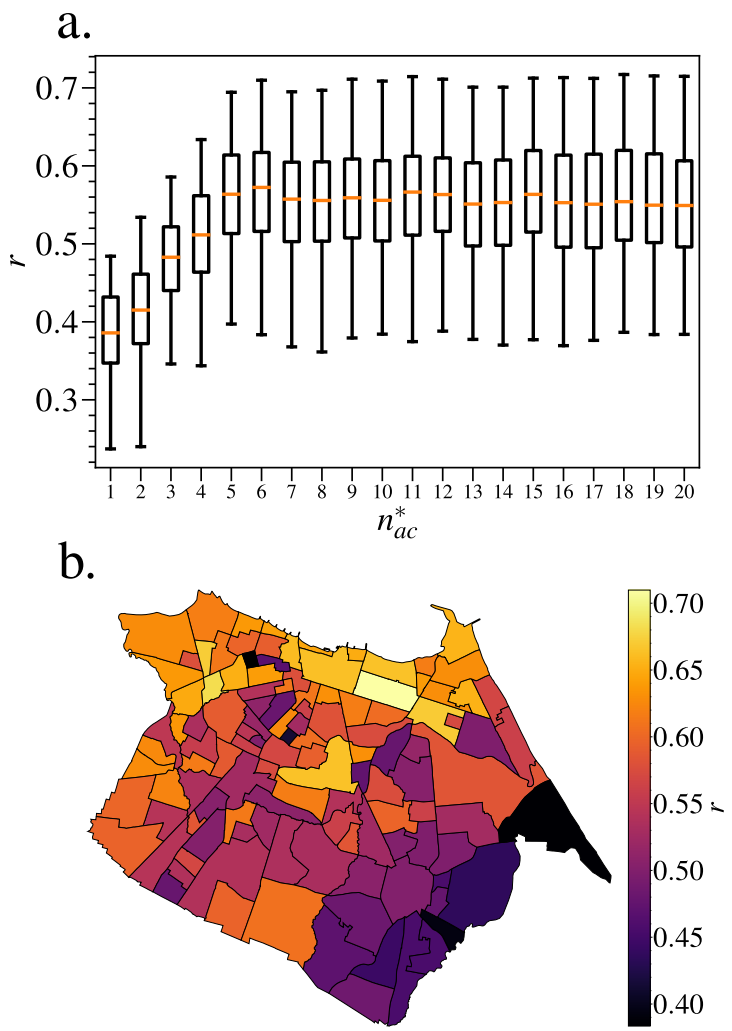

Fig. SI-1. **Analysis of the best threshold** n^*_{ac}. In (a), the boxplot shows the distribution of the Pearson correlation coefficient r versus the number of accumulated reported cases n_{ac}. The median and the 25th and 75th percentiles of each distribution are represented by the orange lines and by the limits of the boxes, respectively. We choose $n^*_{ac} = 6$ as the best threshold value because it is the value that establishes a plateau and maximizes the median of the r distribution. In (b), the map shows the spatial distribution of r for $n^*_{ac} = 6$.