

1 **Human mobility patterns to inform sampling sites for early pathogen detection and routes**
2 **of spread: a network modeling and validation study**

3
4 Prof Andréza L. Alencar^{1#}, PhD, Maria Célia L. S. Cunha^{2#}, PhD, Juliane F. Oliveira^{3*#}, PhD,
5 Adriano O. Vasconcelos², PhD, Gerson G. Cunha², PhD, Ray B. Miranda², BSc, Fábio M. H. S.
6 Filho⁴, BSc, Corbiniano Silva², PhD, Prof Ricardo Khouri⁵, PhD, Thiago Cerqueira-Silva³, PhD,
7 Prof Luiz Landau², PhD, Prof Manoel Barral-Netto^{3,5&}, PhD and Pablo Ivan P. Ramos^{3&}, PhD
8

9 ¹Department of Computer Science, Federal Rural University of Pernambuco; Recife, 52171-900,
10 Brazil.

11 ²Luiz Coimbra Institute of Graduate and Engineering Research (COPPE), Federal University of
12 Rio de Janeiro (UFRJ); Rio de Janeiro, 21941-859, Brazil.

13 ³Center for Data and Knowledge Integration for Health (CIDACS), Gonçalo Moniz Institute,
14 Oswaldo Cruz Foundation (FIOCRUZ); Salvador, 41745-715, Brazil.

15 ⁴Rondonia Oswaldo Cruz Foundation, Oswaldo Cruz Foundation (FIOCRUZ); Porto Velho,
16 76812-245, Brazil.

17 ⁵Medicine and Precision Public Health Laboratory (MeSP²), Gonçalo Moniz Institute, Oswaldo
18 Cruz Foundation (FIOCRUZ); Salvador, 40296-710, Brazil.

19
20 *Corresponding author: juliane.oliveira@fiocruz.br.

21 #These authors contributed equally to this work and are co-first authors.

22 &These authors contributed equally to this work and are co-senior authors.

23
24
25
26
27
28
29
30
31
32

1 **Abstract**

2 **Background:** Detecting and foreseeing pathogen dispersion is crucial in preventing widespread
3 disease transmission. Human mobility is a critical issue in human transmission of infectious
4 agents. Through a mobility data-driven approach, we determined municipalities in Brazil that
5 could make up an advanced sentinel network, allowing for early detection of circulating
6 pathogens and their associated transmission routes.

7 **Methods:** We compiled a comprehensive dataset on intercity mobility spanning air, road, and
8 waterway transport, and constructed a graph-based representation of Brazil's mobility network.
9 The Ford-Fulkerson algorithm, coupled with centrality measures, were employed to rank cities
10 according to their suitability as sentinel hubs.

11 **Findings:** Our results disentangle the complex transportation network of Brazil, with flights
12 alone transporting 79.9 million (CI 58.3 to 10.1 million) passengers annually during 2017-22,
13 seasonal peaks occurring in late spring and summer, and roadways with a maximum capacity of
14 78.3 million passengers weekly. We ranked the 5,570 Brazilian cities to offer flexibility in
15 prioritizing locations for early pathogen detection through clinical sample collection. Our
16 findings are validated by epidemiological and genetic data independently collected during the
17 SARS-CoV-2 pandemic period. The mobility-based spread model defined here was able to
18 recapitulate the actual dissemination patterns observed during the pandemic. By providing
19 essential clues for effective pathogen surveillance, our results have the potential to inform public
20 health policy and improve future pandemic response efforts.

21 **Interpretation:** Our results unlock the potential of designing country-wide clinical sample
22 collection networks using data-informed approaches, an innovative practice that can improve
23 current surveillance systems.

24 **Funding:** Rockefeller Foundation grant 2023-PPI-007 awarded to MB-N.

25
26
27
28
29
30
31

Research in context

Evidence before this study

We searched PubMed on Jun 1, 2023, without language or date restrictions, for the following query: ("mobility network*" OR "transport* network*" OR "sentinel network*" OR "surveillance network*") AND "model*" AND "surveillance". The 469 search results were systematically evaluated, and we identified seven original research studies that applied modeling-based approaches to inform the placement, design, or layout of surveillance/sentinel networks. Of these seven studies, four aimed at optimizing the layout of networks for the monitoring of influenza-like illnesses (ILI), while the others aimed at detecting problems arising from the use of medicines based on pharmacy surveillance; detecting the reporting of common acute conditions through a sentinel network of general practitioners; and optimizing the surveillance strategy for plant pests (*S. noctilio*). Most studies employed maximum coverage algorithms that aim to maximize the protected population. Only a single study incorporated mobility patterns to inform the planning of site placement. Studies that involved ILI sentinel networks were geographically restricted to two United States states (Iowa and Texas), and only one study performed a comprehensive whole of United States modeling.

Added value of this study

Despite the urgent need to improve the capacity and timeliness of clinical sample collection for public health surveillance, very few studies have tackled the design problem for optimal placement of these sampling sites, and even fewer have used large-scale mobility data to inform these design choices in an epidemiologically-relevant way. Our work contributes to this challenge by leveraging airline/roadway/fluvial mobility data for Brazil that, converted into a graph-based representation and using network metrics, allowed us to pinpoint an optimal layout strategy that could improve the current flu surveillance network of this country. Using data collected during the COVID-19 pandemic, we validated the transmission routes and pathways of SARS-CoV-2 spread, confirming that the mobility data-informed spread scenarios recapitulated the actual dissemination of the virus.

Implications of all the available evidence

Mobility data, coupled with network-centered approaches, can complement the identification of strategic locations for early pathogen detection and spread routes.

1 **Background**

2 Detecting emerging pathogens with pandemic potential represents a significant challenge for
3 health authorities and the scientific community, requiring a deep understanding of various
4 factors, including ecological, evolutionary, climatic, and human behavior¹⁻⁶. In particular,
5 pinpointing specific locations suitable for early pathogen detection and functioning as critical
6 hubs for spread is both critical and complex, and should be a critical feature in next-generation
7 surveillance systems⁷.

8
9 Increased human mobility has significantly boosted the risk of new pandemics^{6,8-10}. Today's
10 highly connected world is evidenced by the extensive global transportation network that summed
11 over 38.9 million flights worldwide in 2019¹¹. Additionally, there were around 1.39 million
12 interstate bus trips in 2019¹². Several studies have highlighted the impact of human movement on
13 the spread of infectious diseases, now facilitated by the wealth of available data⁸⁻¹⁰.

14
15 Established strategies for anticipating the emergence of a new pathogen involve monitoring
16 clinical and epidemiological data while prioritizing sampling in areas with a high risk of
17 spillover. However, persistent gaps exist in determining optimal study designs, data collection
18 and sharing methods, and sampling intensity to reliably and timely detect signs of pathogen
19 emergence^{6,13-15}.

20
21 Many countries have created sentinel networks for targeted disease surveillance, with great focus
22 on influenza following the influenza A(H1N1) pandemic of 2009. In Brazil, a surveillance
23 system for flu syndrome relying upon sentinel units was launched nationwide in 2000.
24 Mandatory notification of severe acute respiratory syndrome for hospitalized and deceased cases
25 was implemented in response to the 2009 pandemic. However, akin to many countries,
26 significant challenges limit the representativity of this strategy, including inadequate geographic
27 coverage of sentinel units and *ad hoc* positioning based on existing infrastructure, budget, or
28 mere convenience^{16,17}.

29
30 This study aims to address the challenges of detecting and tracking the spread of infectious
31 diseases at initial stages in a country with continental proportions. This objective is realized by
32 leveraging mobility data and computational algorithms to identify optimal geographic placement

1 of sampling sites—essentially creating a sentinel network with a data-driven and
2 epidemiologically relevant design. By doing so, pathogens known to constitute a threat to public
3 health can be detected early on, and preemptive actions can be taken before widespread
4 transmission occurs. The integration of mobility data into pandemic surveillance offers a fresh
5 perspective on early warning systems, effectively incorporating epidemiological intelligence into
6 the design and planning process.

7

8 **Results**

9 *Spatio-temporal analysis of domestic air transportation*

10 We started by collecting and curating a large-scale dataset for Brazil's transportation network.
11 From 2017 to 2022, 79.9 million passengers flew annually, representing an average of 39.3% of
12 Brazil's population. Seasonal peaks in air traffic data occur in late spring and summer. During
13 this period, October counted an average of 7.14 million passengers (8.9% of yearly average
14 flights), 7.24 million passengers in November (9.1%), and 7.96 million in December (10.0%),
15 with January reaching a maximum of 8.33 million passengers (10.4%). A non-seasonal peak of
16 around 6.94 million passengers is observed in July, representing 8.7% of the yearly passenger
17 flow. However, strict mobility measures impacted these counts in 2020 and 2021. In these years,
18 48.7 and 31.1 million passengers flew, respectively, declines that impacted all 27 federative units
19 of the country.

20

21 Understanding the distribution and dynamics of air transportation infrastructure is crucial for
22 evaluating the potential impact on the spread of infectious diseases, and municipalities with
23 operational airports serve as key nodes in the nationwide transportation network. From 2017 to
24 2023, Brazil averaged 159 such municipalities (out of 5570), with a clear upward trend in the
25 number of municipalities with active flights despite the pandemic-induced decrease in passenger
26 numbers. Most states had fewer than five municipalities with active airports (**Supplementary**
27 **Figure 1**), with notable exceptions including Pará, the state holding the largest number of
28 municipalities with active airports in 2023 (17 airports).

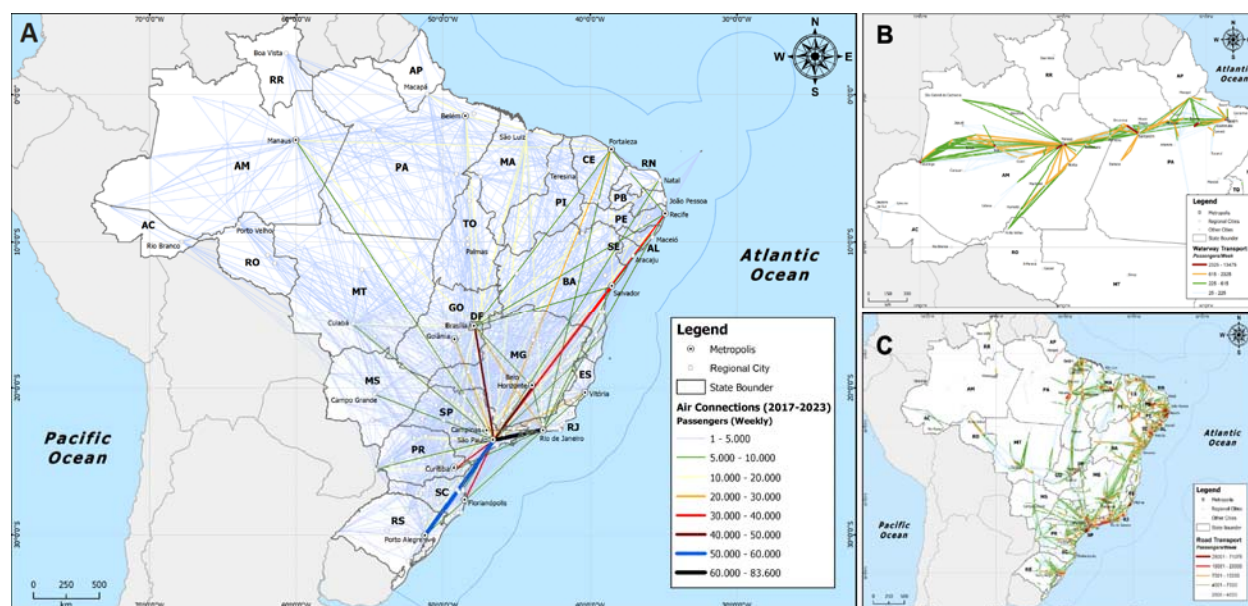
29

30 The Southeastern region of Brazil is a major hub for air travel, hosting an average of 2.4 million
31 incoming passengers (51.3% of all arrivals) and outbound travelers (52.2% of departing
32 passengers) during the examined period. The South follows with 916833 (19.4%)/904268

1 (19.1%) incoming/departing passengers, respectively. Subsequently, the Central-West registered
2 681116 (14.4%)/668255 (14.1%) incoming/departing passengers, the Northeast region had
3 483685 (10.2%)/473309 (10.0%) incoming/departing passengers, and the North region witnessed
4 227430 (4.8%)/222985 (4.7%) incoming/departing passengers (see **Supplementary Figure 2**
5 and **Supplementary Figure 3**).

6
7 Notably, the North and Northeast regions have limited direct connections with the Southern
8 region of Brazil and with each other. Air mobility between these regions may be facilitated by
9 the Southeast and Central-West regions (**Supplementary Figure 3**). Conversely, the Southern
10 region is primarily connected to the Southeast region, with a more fragile connection to the
11 Central-West region. **Figure 1** visualizes the average passenger mobility in air transport,
12 highlighting the influence of states within these regions.

13



14

15 **Figure 1 - Brazil's mobility patterns – airway, road, and waterways.**

16 A) Average passenger volume in the air mobility network from 2017 to 2023. The average
17 frequency of transport by waterways and roads are shown in panels B and C, respectively. Two-
18 letter state abbreviations are as follows: AC, Acre; AL, Alagoas; AP, Amapá; AM, Amazonas;
19 BA, Bahia; CE, Ceará; DF, Distrito Federal; ES, Espírito Santo; GO, Goiás; MA, Maranhão;
20 MT, Mato Grosso; MS, Mato Grosso do Sul; MG, Minas Gerais; PA, Pará; PB, Paraíba; PR,
21 Paraná; PE, Pernambuco; PI, Piauí; RJ, Rio de Janeiro; RN, Rio Grande do Norte; RS, Rio

1 Grande do Sul; RO, Rondônia; RR, Roraima; SC, Santa Catarina; SP, São Paulo; SE, Sergipe;
2 TO, Tocantins.

3

4 *Road and fluvial intercity mobility landscape*

5 We complemented the airline mobility data, which offers monthly totals of passengers flown,
6 with an additional mobility layer that enables the estimation of road/fluvial flows between cities.
7 According to the available data, the country exhibits a weekly mobility capacity of around 78.6
8 million passengers using these transportation modes, representing 38.7% of the Brazilian
9 population.

10

11 Water transport constitutes only 258581 (0.3%) of passenger mobility capacity and is more
12 common in the Northern region of Brazil. In the Amazon state, a substantial 71.4% of mobility
13 capacity relies on river transportation, surpassing road transport. Moreover, this mode of
14 transport forms robust connections, with 99.5% of the link between the states of Amazonas and
15 Pará and 41.3% of the connection between Amazonas and Rondônia facilitated by river
16 transport. Notably, there is a significant water transport connection between the states of Alagoas
17 and Sergipe in Northeastern Brazil, accounting for 12.5% of connectivity. Fluvial transport is
18 also a minority presence in other areas of Brazil, with connections amounting to less than 5%. A
19 representation of passengers' mobility capacity for both roads and waterways in Brazil is shown
20 in **Figure 1**.

21

22 The regional influence on the road and fluvial network also differs from that of air transport. The
23 Northeast region holds the largest share of mobility capacity for passengers at 40.2%, with the
24 Southeast at 28.9%, South at 16.9%, Central-West at 7.6%, and North at 6.1% (**Supplementary**
25 **Figure 3**).

26

27 Similar to the air transportation network, the North and Northeast regions have limited direct
28 connections with the Southern part of Brazil but have more significant connections among
29 themselves (**Supplementary Figure 3**). The Southern region has a more distributed connection
30 with the Southeast and Central-West regions, and the Central-West with the North and Northeast
31 regions compared to the air mobility network (**Supplementary Figure 3**).

32

1 ***Estimating pathogen pathways and sentinel hubs for early pathogen detection***

2 The previously identified mobility patterns allow for the estimation of routes for the spread of
3 infectious diseases by leveraging network metrics such as the betweenness index (BI)
4 (**Supplementary Text 2**). In a nationwide ranking based on the BI, we identified 1391 cities
5 with BI scores equal to or above the BI upper quartile. Using the Ford-Fulkerson algorithm, we
6 derived 7746479 paths initiating from a node in each Brazilian city and concluding at each
7 distinct city with a high BI score. This approach facilitated categorizing and ranking the primary
8 routes for pathogen transmission, revealing strategic cities for close monitoring that could
9 function as sentinel sampling sites in each state (see Data Availability section).

10

11 ***Disentangling transmission routes from key Brazilian cities***

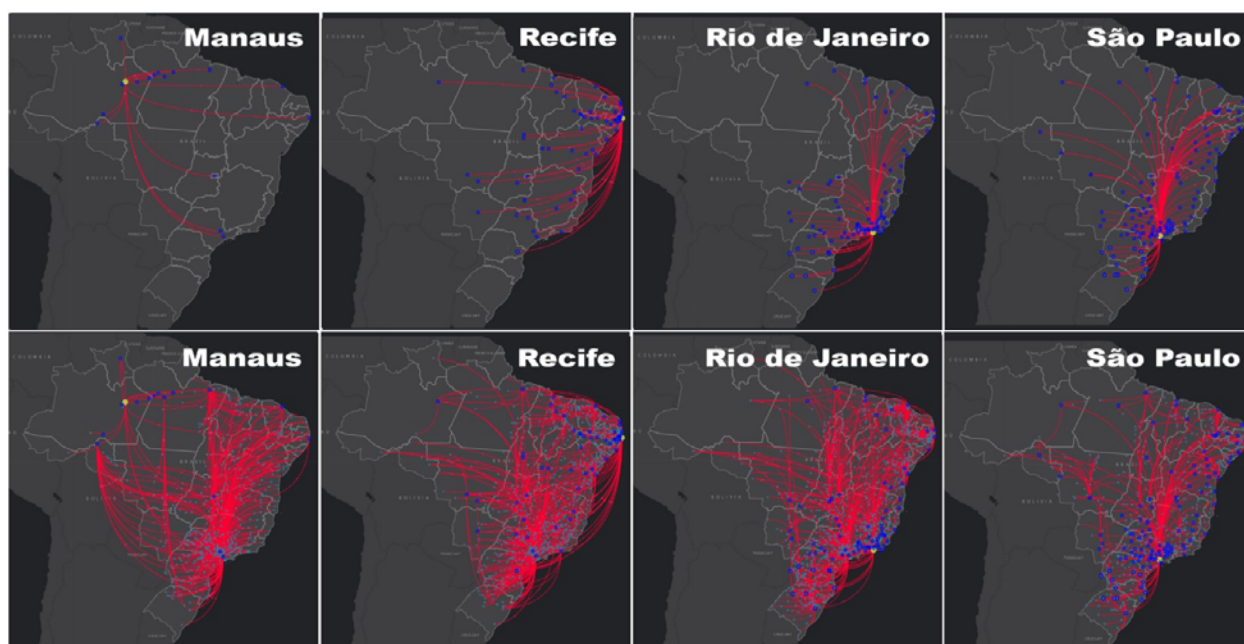
12 We next simulated the emergence of a pathogen initially seeded in four key cities: Manaus (the
13 most populated city in the Amazon region with a population of 2.06 million), Recife (a tourist
14 city in the Northeast with 1.48 million people), Rio de Janeiro and São Paulo, the two largest
15 cities in the country with populations of 6.75 million and 12.3 million, respectively, jointly
16 forming the Megalopolis of the Brazilian Southeast. **Figure 2A** shows the first-step cities in this
17 analysis. Among all paths starting in Manaus and leading to the 1391 cities according to the BI
18 score, 53.6% have São Paulo (city of São Paulo and neighboring municipality Campinas) as their
19 initial destination. This indicates that São Paulo serves as the central primary transmission hub
20 following an emergency in Manaus. Brasília in the Federal District (21.9%), cities in the state of
21 Pará (10.2%), Rondônia (7.8%), Roraima (4.8%), Ceará (0.8%), and Pernambuco (0.5%) are the
22 other most likely states to be affected by a spread emerging in Manaus.

23

24 Recife and Rio de Janeiro exhibit a wider spread potential to other states, and São Paulo plays a
25 crucial role in the subsequent spread from both Rio de Janeiro and Recife, representing 28.7%
26 and 39.6%, respectively, of the first-step cities in the state of São Paulo. As shown in **Figure 2B**,
27 further spread to the second step of the paths would already lead to extensive transmission across
28 the country.

29

30



1
2 **Figure 2 - The spread of a pathogen first detected in Manaus/AM, Recife/PE, Rio de**
3 **Janeiro/RJ, and São Paulo/SP.**

4 The maps in the first row depict the initial propagation range for each path originating in the
5 selected city (first-step cities). In the second row, the extension of these paths for second-step
6 cities are shown. In both panels, yellow dots denote the origin, dark blue dots represent first-step
7 cities, and light blue dots represent second-step cities.

8

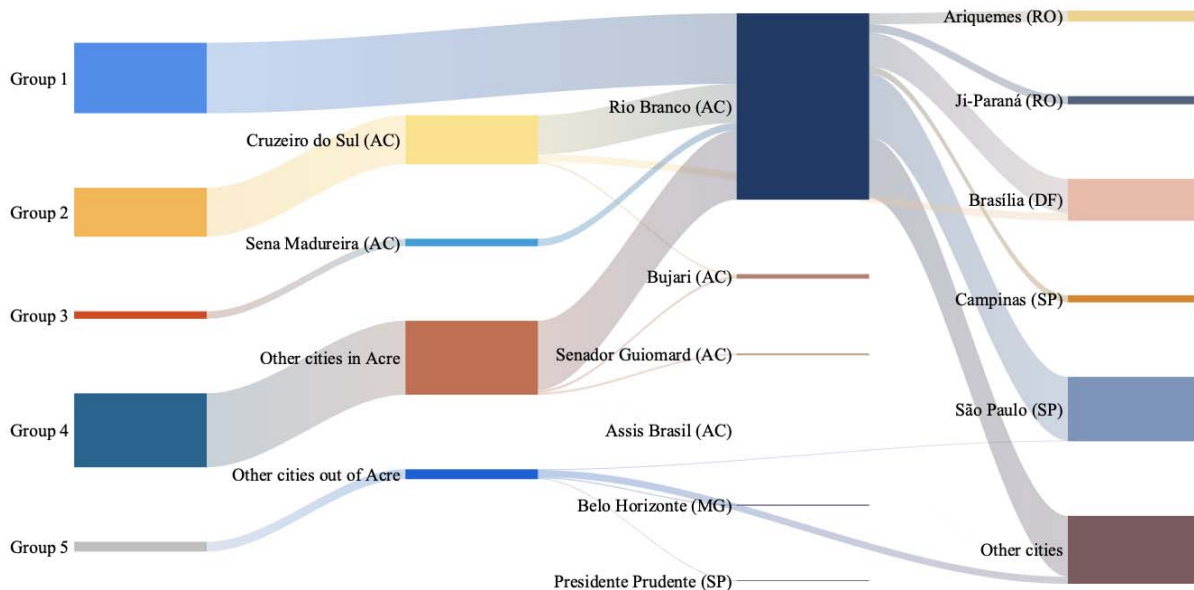
9 *Ranking early warning detection hubs and state gateway cities*

10 Analyzing all pathways starting from cities in each state enables the creation of a ranking that
11 assesses the likelihood of early pathogen detection capacity of each city. This ranking further
12 facilitates the identification of cities according to their significance as mobility gateways to other
13 states. Mobility gateways are defined as cities for which the first step in the most likely paths
14 originating from them leads to cities in other states. Data for the states of Acre, Amazonas, and
15 Rio de Janeiro are presented in **Supplementary Table 1**.

16

17 According to the BI score, pre-selected cities in Acre are Rio Branco (the state capital), Cruzeiro
18 do Sul, and Sena Madureira. In the event of a pathogen emergency in a city in Acre that goes
19 undetected, Rio Branco and Cruzeiro do Sul emerge as the most probable cities for pathogen
20 detection using the sentinel approach. Approximately 33.5% and 23.2% of paths starting in Acre
21 use these cities as the first step, respectively. **Figure 3** provides a visual representation of the

1 pathways from cities in Acre to the 1391 selected semi-hubs across the country. Moreover, the
2 list of gateway cities features Rio Branco and Sena Madureira as the key priority points for
3 mobility to other cities outside of Acre.



4
5 **Figure 3 - Transmission routes for cities in the state of Acre, North Brazil.**

6 Groups 1, 2 and 3 represent cities in Acre, with Rio Branco, Cruzeiro do Sul and Sena Madureira
7 as their respective first-step cities. Group 4 comprises cities with the first step being a non-pre-
8 selected city according to the BI in Acre. Group 5 includes cities with their first step being cities
9 with high BI scores in other states.

10
11 Similar to Acre, the priority ranking for early detection cities in the state of Amazonas could not
12 solely be determined by the BI score selection. In this region, the capital city Manaus, followed
13 by Tefé, emerge as the most likely candidates for early pathogen detection. These cities account
14 for 37.7% and 8.6%, respectively, of paths leading to them as the initial step.

15
16 Amazonas has 66 cities, with 16 serving as gateways for transmission to other states. These
17 gateway cities can be sorted into two categories according to their BI: The first group,
18 comprising gateway cities with high BI exhibit a broader reach, allowing passenger flows toward
19 more distant states such as São Paulo, the Federal District, Ceará, Pernambuco, as well as
20 neighboring states like Rondônia, Roraima, and Pará. In contrast, gateway cities with low BI

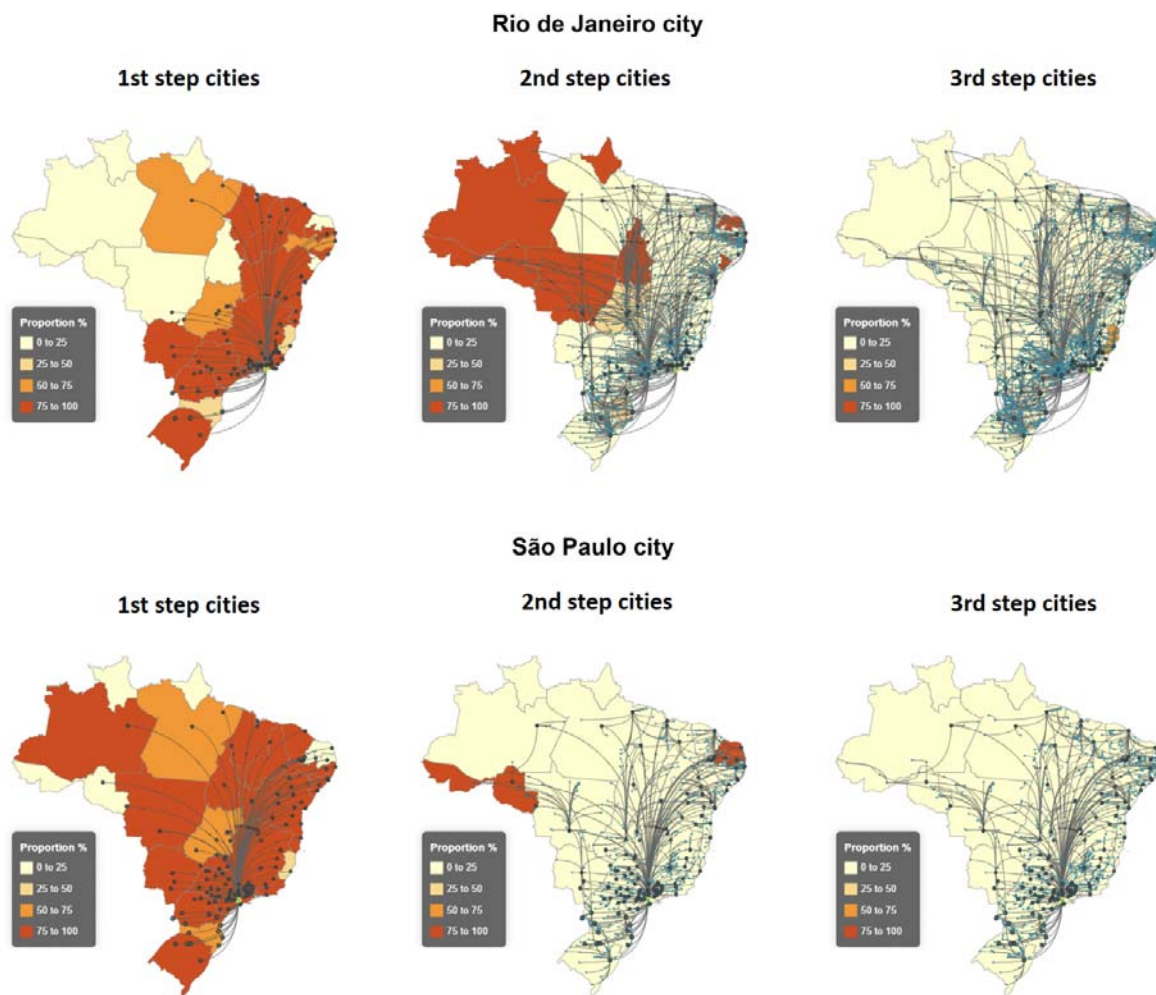
1 scores (falling in the second group) primarily impact neighboring states, including Rondônia,
2 Acre, Roraima, and Pará.

3
4 In the case of Rio de Janeiro state, the data reveals a pattern where pre-selected cities according
5 to the BI dominate as the top five candidates for early pathogen detection. The state capital, Rio
6 de Janeiro city, represents a substantial portion, accounting for 40.4% of paths originating from
7 these cities as the initial step. Following a similar analysis as that conducted for Amazonas, we
8 observe that gateway cities in Rio de Janeiro, that satisfy our pre-selection criterion, contribute to
9 a broader transmission spanning 19 states out of the 27. In contrast, gateway cities in Rio de
10 Janeiro that are not pre-selected tend to have a more localized impact, reaching nearby states
11 such as São Paulo, Espírito Santo, and Minas Gerais, as well as more distant states like Tocantins
12 and Bahia.

13
14 ***Validation with the initial entry period of SARS-CoV-2 in Brazil (February-April, 2020)***

15 During the early stage of the COVID-19 pandemic in Brazil, the first reported cases emerged in
16 the cities of São Paulo and Rio de Janeiro on February 26, 2020 and March 5, 2020, respectively.
17 In this period, a total of 44 and 13 cases of SARS-CoV-2 infection were reported in each city,
18 respectively. Another 19 states had reported cases, although they did not exceed five cases and
19 were mostly related to imported cases at that point. In **Figure 4**, we show the relationship
20 between the proportion of cases in cities in the first, second, and third steps relative to the total
21 number of reported cases in each state until April 1, 2020.

22



1
2 **Figure 4 - Epidemiological validation of pathogen transmission pathways.**
3 Spatial distribution of early-stage (up to April 1, 2020) cities affected by the spread of SARS-
4 CoV-2 originating in the cities of Rio de Janeiro and São Paulo. Brazilian states are colored
5 based on the proportion (%) of COVID-19 cases reported in cities in the first-, second-, and
6 third-steps relative to the total number of reported cases in each state. Lines depict the paths in
7 first step-cities, second-step cities, and third-step cities.

8
9 **Figure 4** evidences that the 83 first-step cities identified for Rio de Janeiro municipality
10 recapitulate the actual pattern of initial SARS-CoV-2 dissemination during the studied period,
11 with 66.0% of the cases reported in these cities. Most of these first-step cities locate along the
12 Brazilian shoreline (**Figure 4**). Then, in second-step cities the disease spread towards the North
13 and Mid-West regions, concentrating less connected cities that require additional intermediate
14 connections to be reached.

1
 2 On the other hand, São Paulo, one of Latin America's largest cities, presents with a much higher
 3 transmission capacity across the country, with first-step cities distributed in more states
 4 compared to Rio de Janeiro (**Figure 4**). Similarly, 83·8% of cases were identified in first step
 5 cities. The states of Rondônia and Acre in the North and Paraíba and Rio Grande do Norte in the
 6 Northeast are the states with a lower proportion of cases in first-step cities on the paths of São
 7 Paulo, in agreement with the strength of connection between these states and the state of São
 8 Paulo. Finally, the following steps (2nd and 3rd cities) show a pattern consistent with the observed
 9 spatial spread (**Figure 4**).

10
 11 ***Validation with the initial entry period of the Gamma (P.1) variant in Amazonas (January-***
 12 ***March, 2021)***

13 SARS-CoV-2 sequencing data produced during the entry of the Gamma (P.1) variant of concern,
 14 which emerged in December, 2020 in the Amazonas state¹⁸, was also used to validate the
 15 sentinel approach. By January 6, 2021, a total of 133 cases were detected in Manaus, with a few
 16 sporadic cases detected in other locations (fewer than 13). Table 1 highlights that most cases
 17 concentrated in cities at the first step of paths originating from Manaus during the period before
 18 February 1, 2021. As the Gamma variant further spread, a larger proportion of cases emerged in
 19 cities at the second step. Confirming the data shown for the initial entry of SARS-CoV-2 in
 20 Brazil, this additional, real-world result on the spatial distribution of this variant is consistent
 21 with the mobility patterns illustrated in **Figure 2**.

22
 23 **Table 1** - Case distribution in paths originating from Manaus during the early-stage
 24 dissemination of the Gamma (P.1) variant in Brazil.

Period	Total detected Gamma cases	N. of cases detected in 1 st step cities (% of total detected cases)	N. of cases detected in 2 nd step cities (% of total detected cases)	N. of cases detected in 3 rd step cities (% of total detected cases)	N. of cases detected in other cities (% of total detected cases)

before January 16,2021	307	224 (73%)	49 (16%)	6 (2%)	28 (9%)
before February 1, 2021	583	305 (52%)	187 (32%)	26 (4%)	65 (11%)
before March 1, 2021	2208	603 (27%)	1075 (49%)	172 (8%)	358 (16%)

1

2

3 Discussion

4 Human mobility is recognized to have a chief role in the management of outbreaks, evidenced by
5 the travel restrictions enacted during COVID-19^{8,19}, and during the 2009 H1N1 pandemic⁹. Here,
6 we unveiled the extensive mobility landscapes of one of the world's largest countries, Brazil,
7 encompassing air, road, and water transportation routes. Key regions and transport modes
8 significantly contributing to Brazil's inter-city connectivity were identified. Our analysis delves
9 into the historical dynamics of air transport availability in Brazilian cities, highlighting seasonal
10 variations and examining the impact of the COVID-19 pandemic on air travel. This temporal
11 variation serves as a valuable lens to comprehend the influence of pathogen spillovers and
12 spread, aligning with findings in various studies that have used mobility data to assess the spatial
13 dissemination of diseases^{8-10,20}.

14

15 In particular, we pursued a ranking strategy leading to a data-informed ordering of cities that
16 could make up a sentinel network for clinical sample collection, enhancing the design of Brazil's
17 currently used flu syndrome surveillance system. Likewise, other works have attempted to
18 address this problem, notably that of Cheng *et al.*, who formulated this as a multi-objective
19 numeric optimization task, enabling simulation scenarios to be compared and providing a
20 framework for optimal site placement¹⁷. However, no real-world application of that approach
21 was attempted. Others have used real-world data to adapt local influenza sentinel networks by
22 relying on the concept of maximally covered population, in which the number and location of
23 sentinel sites were chosen to elicit a configuration that maximizes the targeted population^{16,21}.

1 However, the catalytic power of human mobility on disease spread was not considered in these
2 works.

3
4 On the other hand, the construction of a mobility-driven metapopulation model covering 35
5 states in the United States has been reported²². However, despite common overarching goals, our
6 work presents several methodological differences. While the mobility dataset used by Pei *et al.*
7 focuses on commuting (i.e., journey to work) behavior within and between neighboring counties,
8 we were able to collect a national dataset of transportation patterns covering air/road/fluvial
9 transportation modes. Most importantly, our approach is pathogen-agnostic (i.e., no assumptions
10 around the characteristics of an emerging disease, such as transmission rates, are made, other
11 than that its spread is directly influenced by human flow), while mathematical models such as
12 SIR and extensions thereof²² rely on several premises, changes of which can greatly impact the
13 results of these models. Predicting the exact location and time of emergence of a new pathogen is
14 a challenging task. Pathogens can appear unnoticed even in unusual places and remain
15 undetected for an extended period of time before triggering a major outbreak (as with Zika virus,
16 or the different SARS-CoV-2 variants). Our results consider every Brazilian city as a potential
17 site for a spillover and provide guidance on how to respond when a new pathogen emerges.

18
19 The primary pathogen transmission routes originating at each Brazilian city were identified by
20 applying the Ford-Fulkerson algorithm to Brazilian large-scale mobility data. Additionally, we
21 disclosed strategic cities for early pathogen detection and recognized gateway cities in each state,
22 offering insights into potential widespread transmission. Our results pointed to distinct patterns
23 in Brazil's Central-West and Southeastern regions of Brazil, where economically prosperous
24 cities are elemental in facilitating mobility among the Northern, Southern, and Northeastern
25 regions. The well-established infrastructure for human mobility in these regions, marked by
26 airport expansion and modernization efforts, emphasizes the urgency of a strategic early
27 detection plan in this context. Such an approach is particularly crucial to mitigate the spread of
28 diseases in less affluent Northern and Northeastern areas, where early detection strategies must
29 account for the scarcity of infrastructure for pandemic surveillance, laboratory, and sampling in
30 these regions, compared to the South/Southeast.

31

1 State capitals should not be the sole focus when determining hub cities for primary sampling.
2 Cities within a state may require an intermediate hub for maximal disease spread, and some
3 cities, with direct and robust connections, may function independently from capitals, influencing
4 mobility to other locations. Finally, ranking gateway cities for further mobility provides valuable
5 insights for planning early detection and containment strategies at the state level. Our results
6 strongly agree with epidemiological and genomic data collected during the entry of the gamma
7 (P.1) variant in North Brazil, evidencing the utility of mobility-based sentinel surveillance
8 placement in this continental-sized country^{23–25}.

9
10 This study has limitations: The limited availability of more up-to-date mobility data on roads and
11 waterways, estimated only up to 2016. The collected mobility data does not encompass private
12 transport, and inter-city mobility for cities sharing common borders is either underestimated or
13 poorly recorded. One potential strategy to mitigate this limitation is to treat these border cities as
14 a collective unit when planning the deployment of early-alert and pathogen detection systems.
15 Despite these challenges, our work offers valuable insights that can serve as a catalyst for future
16 improvements in mobility data collection methodologies.

17
18 Beyond the current scope lies the potential inclusion of variables influencing the selection of
19 cities as hubs for early pathogen detection. Factors such as local health priorities and resources,
20 characteristics favoring the emergence of specific pathogens, the availability of comprehensive
21 health services covering all age groups, significant locations where workers have direct contact
22 with animals, and others could be integrated with our ranking list. This integration could enhance
23 the deployment of sentinel locations for early detection and improve predictions of transmission
24 routes.

25 26 **Material and Methods**

27 *Data sources*

28 We collected and curated large-scale mobility data from different administrative sources to
29 enable a comprehensive understanding of inter-city mobility via air, road, and river networks.
30 For validation of the findings, we used epidemiological and genetic circulation data of SARS-
31 CoV-2 within Brazilian municipalities during the Gamma (P.1) period.

32

1 *Mobility data*

2 Inter-city mobility data covering the Brazilian territory was collected from the Brazilian Institute
3 of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística - IBGE) for both road
4 and river networks; air transport data was obtained from the National Civil Aviation Agency
5 (Agência Nacional de Aviação Civil - ANAC)^{26,27}.

6
7 The IBGE database provides the weekly frequency of transport capacity between city pairs,
8 estimated for 2016. The ANAC database contains extensive historical information on regular and
9 non-regular public air transport services operated by domestic and foreign companies, as well as
10 private flights, in Brazil from 2000 onwards. We used data covering the period of 2017-2023
11 regarding the number of passengers on domestic flights to allow portraying the air mobility
12 landscape in the country prior to (i.e., 2017-2019), during (i.e., 2020-2021) and after the
13 COVID-19 pandemic period (i.e., 2022-2023).

14
15 To align the IBGE and ANAC databases, we used the IBGE's data collection methodology to
16 convert monthly passenger numbers in ANAC into weekly counts^{27,28}. In **Supplementary Text 1**
17 the steps for normalization of these data sources are presented.

18
19 ***Validation of the analytical framework using SARS-CoV-2 transmission data***

20 The analytical model was validated two-fold using data captured during the COVID-19
21 pandemic period. First, we collected data from confirmed COVID-19 cases for each Brazilian
22 municipality until May 1, 2020, comprehending the early stage of SARS-CoV-2 entry and spread
23 in the country. The data is openly available and provided by the Brazilian Ministry of Health²⁹.

24
25 This validation using real-world data during the beginning of the spread aimed to confirm
26 whether the mobility-based spread routes, as determined in our simulations, would recapitulate
27 the actual spread pathways that took place in the initial course of COVID-19 transmission
28 between February 26 and March 12, 2020, before community transmission was declared in the
29 country (March 13, 2020)²⁰. Thereafter, transmission rates of the disease were altered due to
30 non-pharmaceutical interventions likely affecting the spread patterns. Then, we evaluated the
31 proportion of cities that reported a case in paths obtained using the Ford-Fulkerson algorithm
32 (described below) until April 1, 2020.

1
2 In a second, independent validation approach, we leveraged SARS-CoV-2 genomic data to
3 examine the rises and spreading patterns in COVID-19 cases following the Gamma (P.1) variant
4 emergence in Manaus (North region), which according to existing literature originated in that
5 region²³⁻²⁵. We considered three entry/spread periods: 1) from January 6 to January 16, 2021; 2)
6 until February 1, 2021; and 3) until March 1, 2021. Combined, these validation strategies aimed
7 to verify whether the analytical framework employed would recapitulate the geographic spread
8 of a newly-emerged pathogen or a sub-variant for which the affected population would be largely
9 immunologically naïve.

10

11 *Construction of the mobility network and estimation of transport flows*

12 The intercity mobility data was used to build a mobility network defined by the graph $G = (V,$
13 $E)$. Here, $V = \{v_1, v_2, \dots, v_n\}$ represents a set of n nodes, where each node $v_i \in V$
14 corresponds to a city. Additionally, $E = \{e_1, e_2, \dots, e_m\}$ represents a set of m edges, where
15 each edge $e_i \in E$ holds the frequency of mobility between a pair of cities as an associated
16 weight. The Ford-Fulkerson method computes the maximum flow that can be sent from a
17 designated source node to a designated sink node through a network of interconnected nodes and
18 directed edges³⁰. **Supplementary Figure 4** outlines the steps involved in the iterative
19 computation of the Ford-Fulkerson algorithm.

20

21 This mobility network, coupled with the Ford-Fulkerson method to estimate flows, were used to
22 infer the most suitable locations for early detection and to track the most likely trajectory of a
23 (newly) emerged pathogen in Brazilian cities. Due to the computational complexity of this task,
24 requiring the evaluation of over 31 million combinations of paths between each pair of cities, we
25 selected a subset of cities based on their network centrality. The Betweenness Index (BI), a
26 global centrality measure, was used as the selection criterion. The cities with BI exceeding the
27 third quartile were sought as destinations for our calculated routes starting in each of the 5570
28 Brazilian cities. The BI is defined as

$$BI(v_i) = \frac{\sum g_{jk}(v_i)}{\sum g_{jk}}$$

29

1 where g represents the number of geodesic paths (shortest paths) between two nodes in the
2 network, and $g_{jk}(v_i)$ represents the geodesics containing the node v_i , with $i \neq j \neq k$ ³¹. We
3 also compared the BI to the proximity index, a secondary centrality metric that is highly
4 correlated with the BI (**Supplementary Text 2**).

5

6 ***Role of the funding source***

7 The funders had no role in the study design, data collection, analysis, interpretation and writing
8 of this report, nor on the decision to submit the paper for publication.

9

10 **Acknowledgments**

11 This study is part of the Alert-Early System of Outbreaks with Pandemic Potential (ÆSOP,
12 <http://aesop.health>), an initiative under development by Brazil's Fundação Oswaldo Cruz
13 (Fiocruz) and the Federal University of Rio de Janeiro with financial support from the
14 Rockefeller Foundation's Health Initiative (Grant 2023-PPI-007 awarded to M-BN). The
15 excellent project management and administrative support from Fundação de Apoio à Fiocruz
16 (FIOTEC) is greatly appreciated.

17

18 **Author contributions**

19 Conceptualization: ALA, MCLSC, JFO, LL, MB-N, PIPR. Methodology: ALA, MCLSC, JFO,
20 AOV, FMHSF, RK, LL, RBM, PIPR. Investigation: ALA, MCLSC, JFO, AOV, PIPR.
21 Visualization: MCLSC, JFO, AOV, GGC, RBM, CS, PIPR. Funding acquisition: MB-N. Project
22 administration: MB-N, PIPR. Supervision: LL, MB-N, PIPR. Writing – original draft: ALA,
23 JFO, PIPR. Writing – review & editing: all authors.

24

25 **Declaration of interests**

26 Authors declare that they have no financial or personal relationships that could be construed as
27 competing interests.

28

29 **Data and materials availability**

30 All data needed to evaluate the conclusions in the paper are presented in the paper and/or the
31 accompanying [GitHub](https://github.com/andrezaite/reproducibility_transportation_hubs-) repository
32 https://github.com/andrezaite/reproducibility_transportation_hubs-

1 [early_warning_surveillance_systems.git](https://github.com/early-warning-surveillance-systems.git). The SARS-CoV-2 genome data used in this study are
2 sourced from GISAID (<http://gisaid.org>), with accession number gisaid_hcov-
3 19_2022_12_1800. A dashboard enabling direct access to state hubs ranking, gateway cities and
4 paths of transmissions is available and described in the aforementioned GitHub repository, under
5 the *visualization* folder.

6

7 **Supplementary Materials**

8 Supplementary Text 1 and 2

9 Supplementary Figures 1 to 4

10 Supplementary Table 1

11

12 **References**

13 1 Nachegea JB, Nsanzimana S, Rawat A, *et al.* Advancing detection and response capacities
14 for emerging and re-emerging pathogens in Africa. *Lancet Infect Dis* 2023; **23**: e185–9.

15 2 Edwards AM, Baric RS, Saphire EO, Ulmer JB. Stopping pandemics before they start:
16 Lessons learned from SARS-CoV-2. *Science* 2022; **375**: 1133–9.

17 3 Taubenberger JK, Kash JC, Morens DM. The 1918 influenza pandemic: 100 years of
18 questions answered and unanswered. *Sci Transl Med* 2019; **11**.
19 DOI:10.1126/scitranslmed.aau5485.

20 4 Jones KE, Patel NG, Levy MA, *et al.* Global trends in emerging infectious diseases.
21 *Nature* 2008; **451**: 990–3.

22 5 Gibb R, Redding DW, Chin KQ, *et al.* Zoonotic host diversity increases in human-
23 dominated ecosystems. *Nature* 2020; **584**: 398–402.

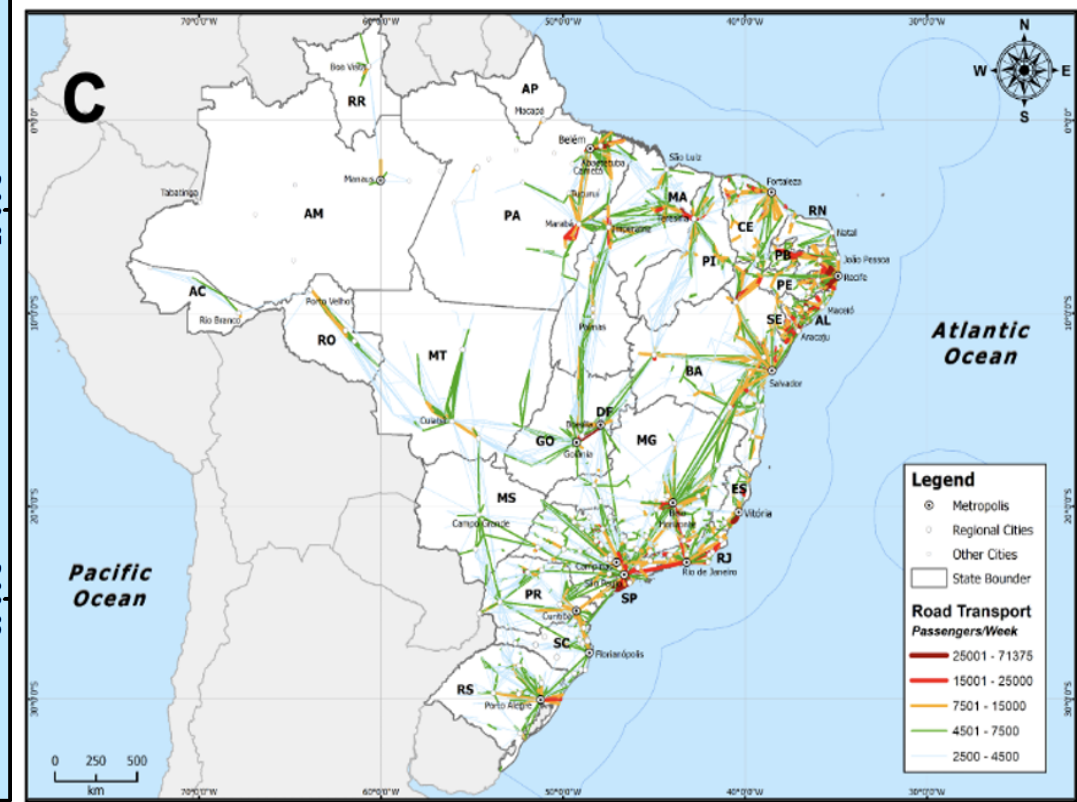
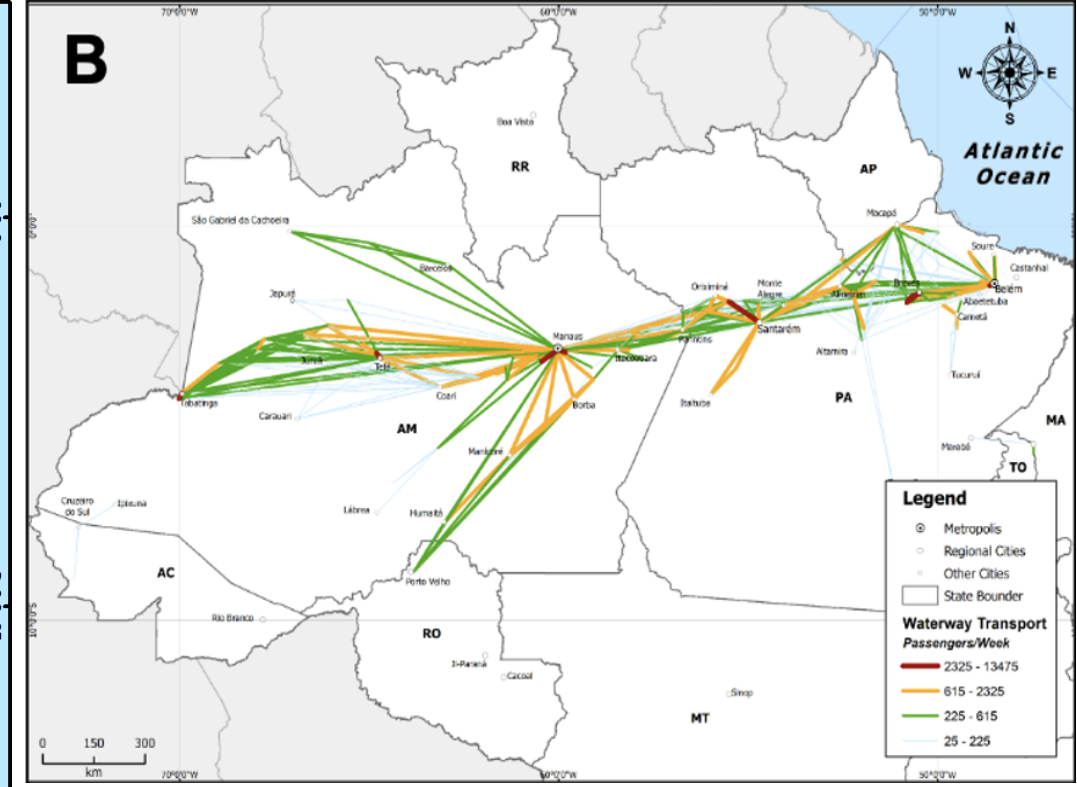
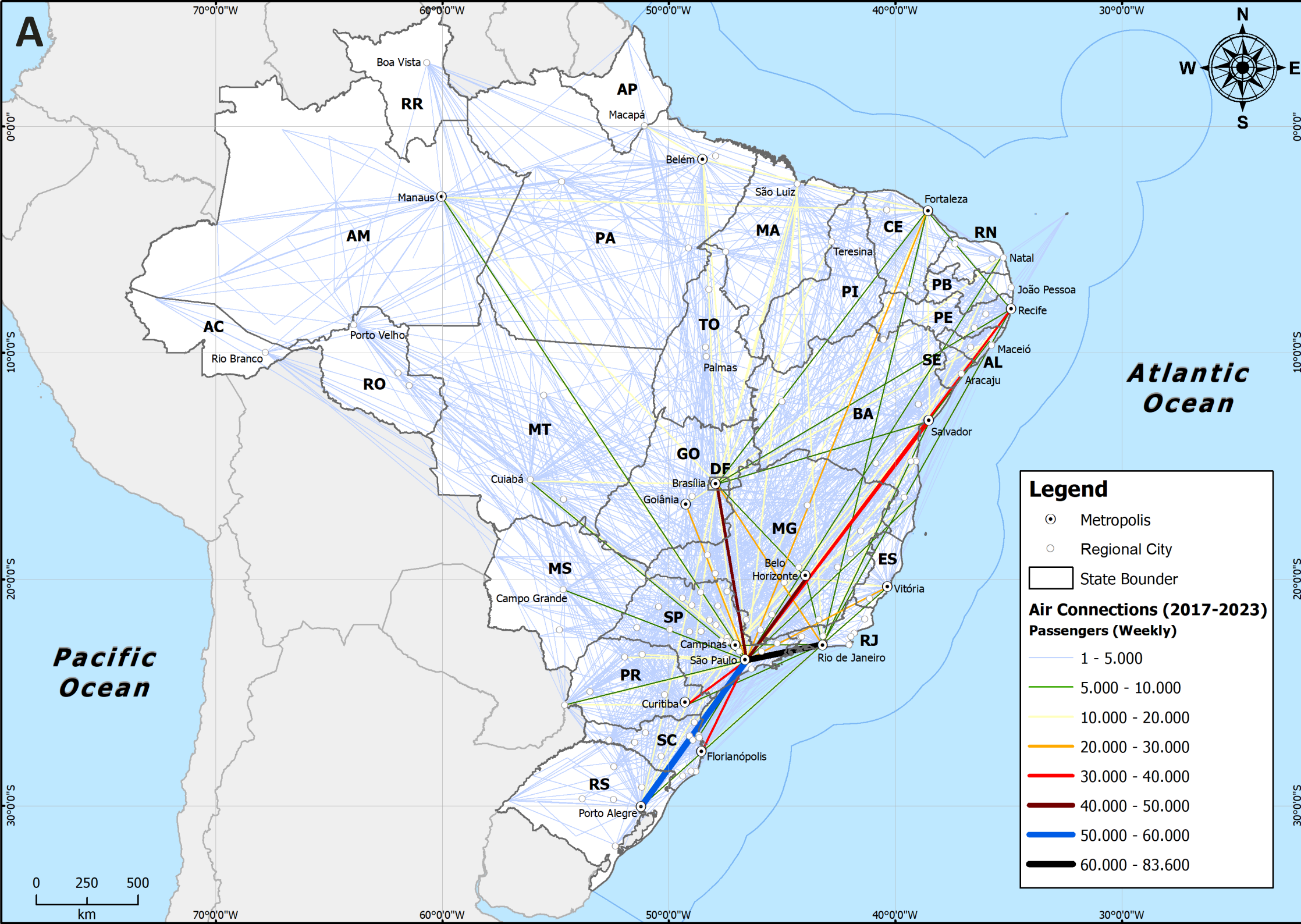
24 6 Kraemer MUG, Cummings DAT, Funk S, *et al.* Reconstruction and prediction of viral
25 disease epidemics. *Epidemiol Infect* 2018; **147**: e34.

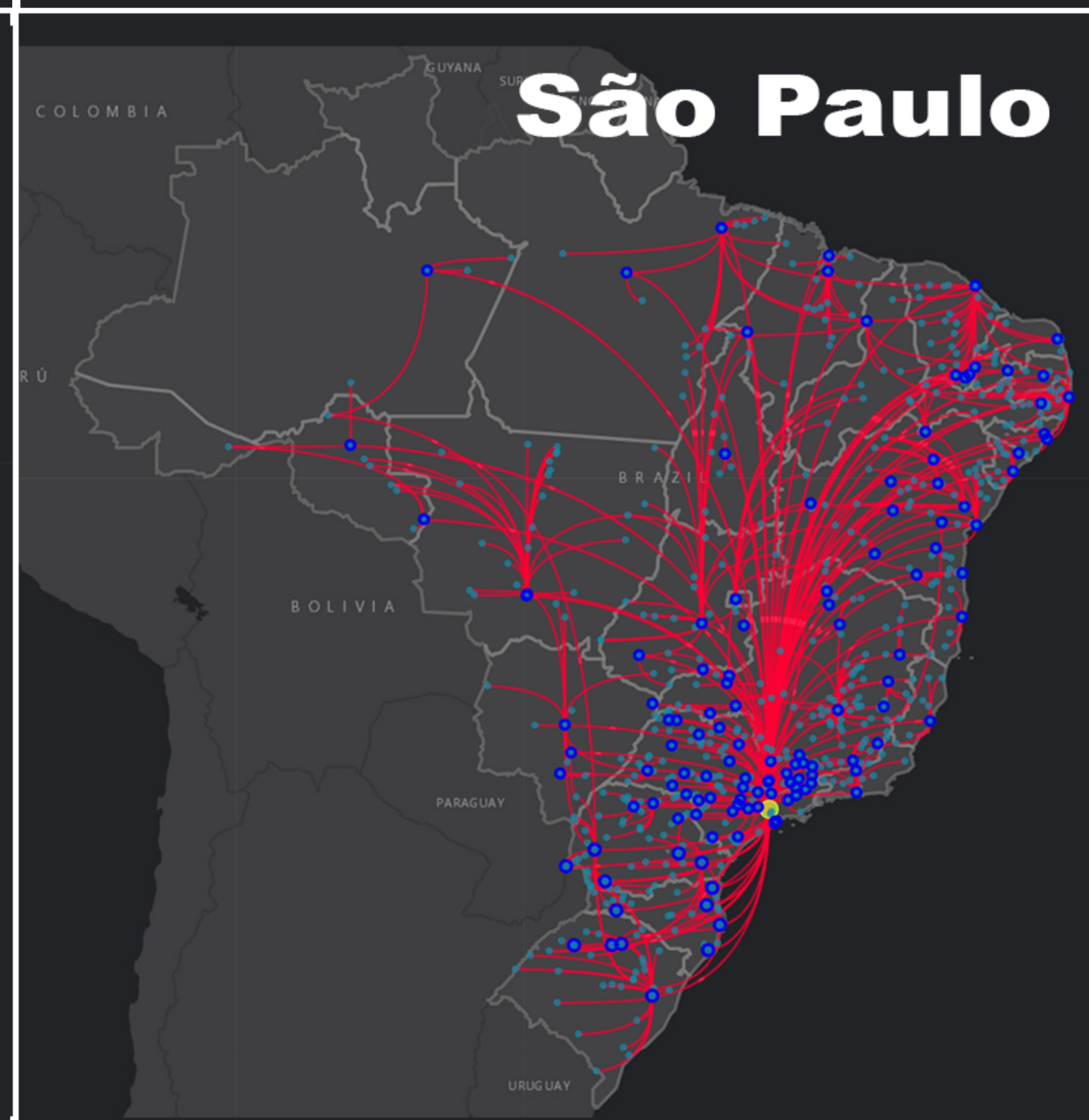
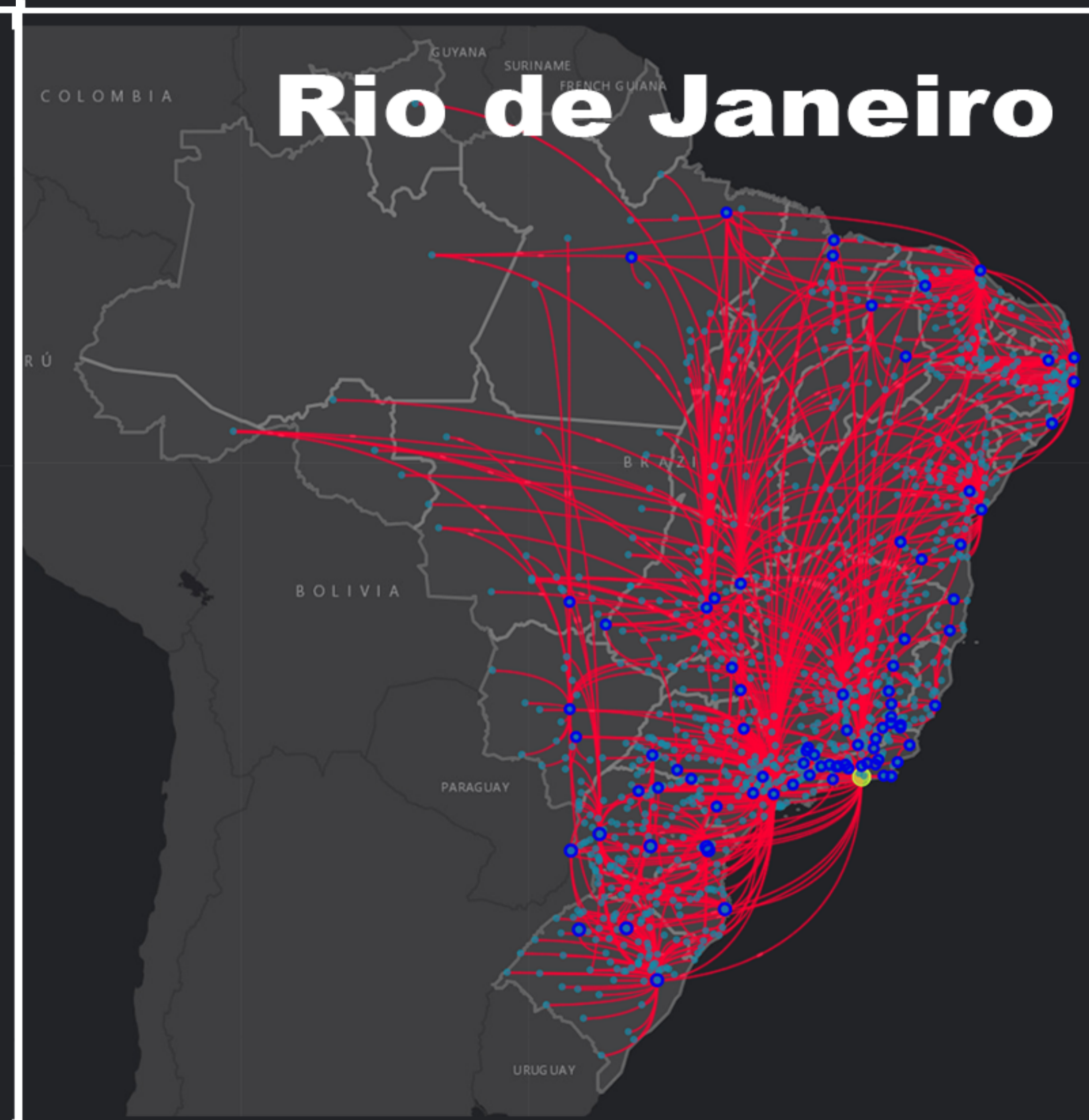
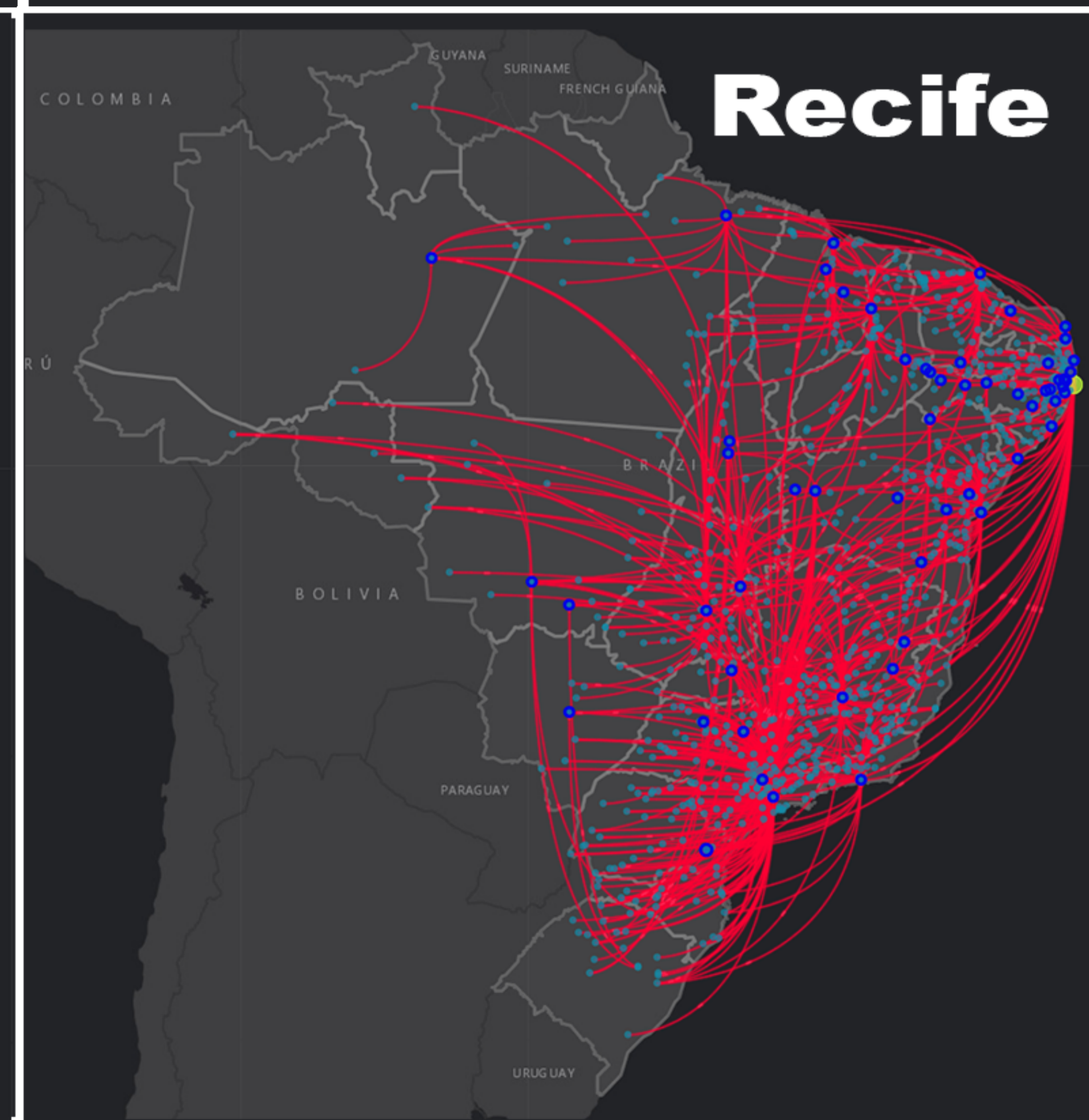
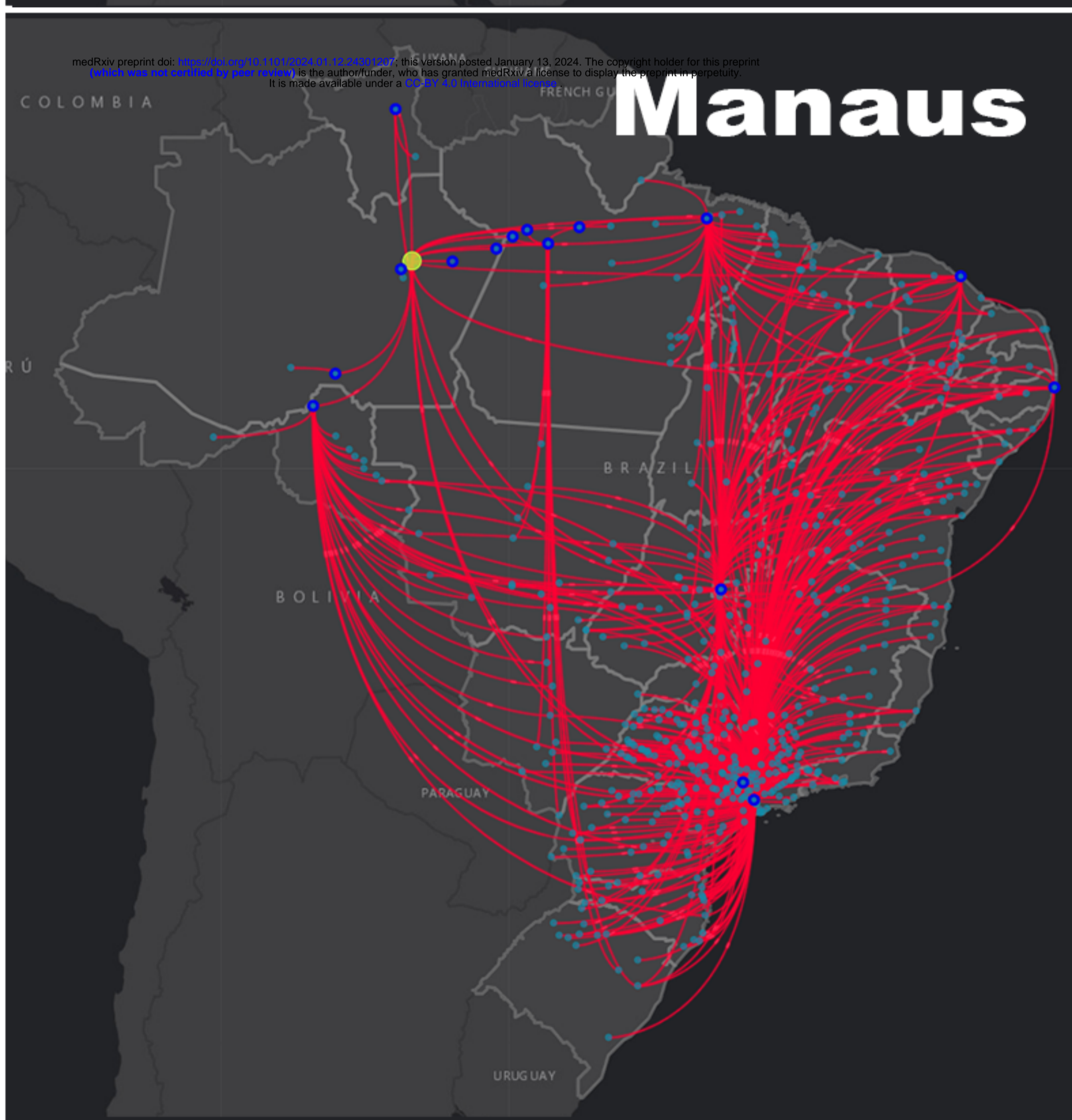
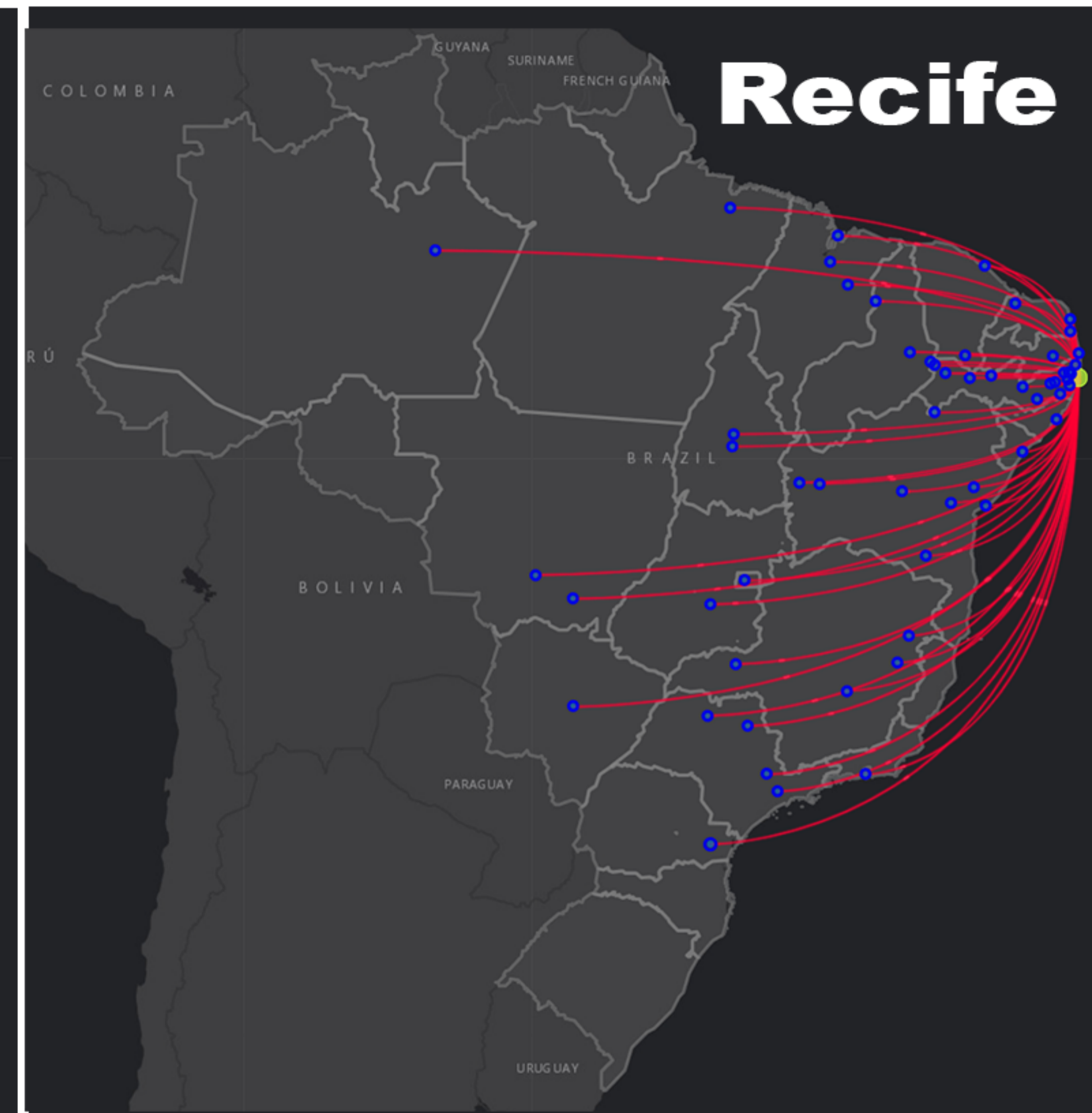
26 7 Ramos PIP, Marcilio I, Bento AI, *et al.* Combining Digital and Molecular Approaches
27 Using Health and Alternate Data Sources in a Next-Generation Surveillance System for
28 Anticipating Outbreaks of Pandemic Potential. *JMIR Public Health Surveill* 2024; **10**:
29 e47673.

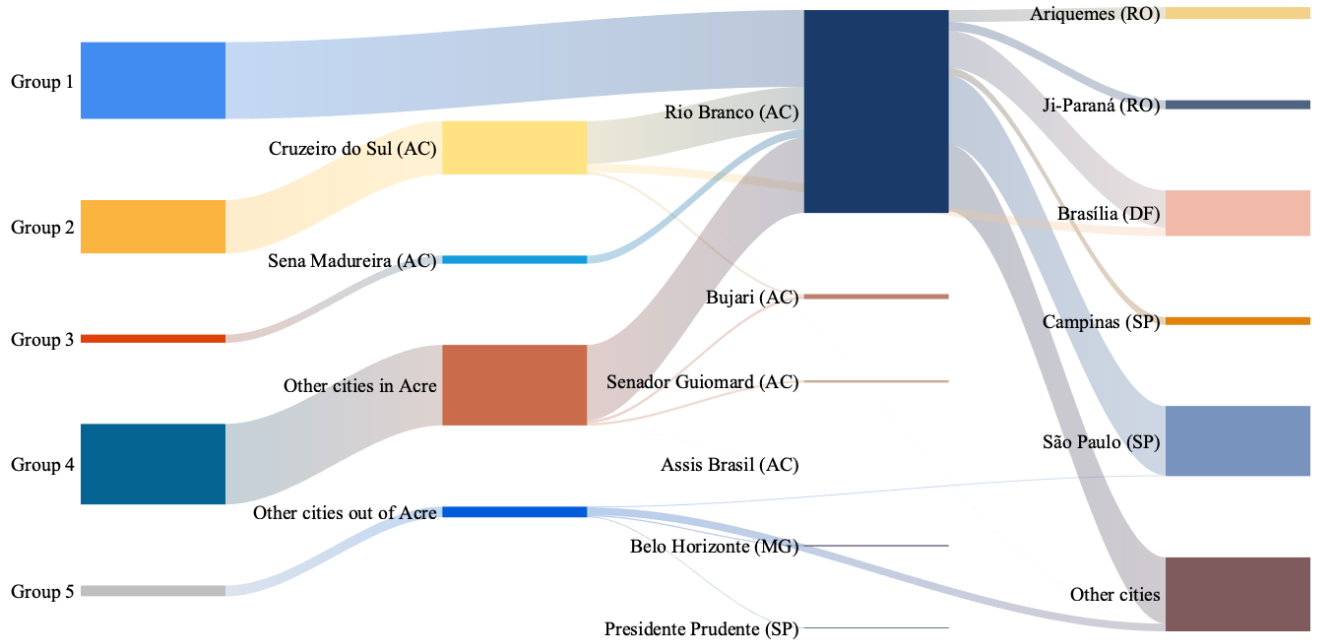
30 8 Mu X, Yeh AG-O, Zhang X. The interplay of spatial spread of COVID-19 and human
31 mobility in the urban system of China during the Chinese New Year. *Environment and*
32 *Planning B: Urban Analytics and City Science* 2021; **48**: 1955–71.

- 1 9 Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A. Human mobility
2 networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE*
3 2011; **6**: e16591.
- 4 10 Peixoto PS, Marcondes D, Peixoto C, Oliva SM. Modeling future spread of infections via
5 mobile geolocation data and population dynamics. An application to COVID-19 in Brazil.
6 *PLoS ONE* 2020; **15**: e0235732.
- 7 11 Gullen A, Plungis J. *Statista. charleston adv* 2013; **15**: 43–7.
- 8 12 Ministério da Infraestrutura. Ministry of Infrastructure, Brazil.
9 <http://portal.infraestrutura.gov.br/> (accessed Dec 20, 2023).
- 10 13 Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: Make outbreak research open
11 access. *Nature* 2015; **518**: 477–9.
- 12 14 Quick J, Loman NJ, Duraffour S, *et al.* Real-time, portable genome sequencing for Ebola
13 surveillance. *Nature* 2016; **530**: 228–32.
- 14 15 Grubaugh ND, Ladner JT, Kraemer MUG, *et al.* Genomic epidemiology reveals multiple
15 introductions of Zika virus into the United States. *Nature* 2017; **546**: 401–5.
- 16 16 Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, Rushton G. Optimizing
17 influenza sentinel surveillance at the state level. *Am J Epidemiol* 2009; **170**: 1300–6.
- 18 17 Cheng Q, Collender PA, Heaney AK, *et al.* The DIOS framework for optimizing
19 infectious disease surveillance: Numerical methods for simulation and multi-objective
20 optimization of surveillance network architectures. *PLoS Comput Biol* 2020; **16**:
21 e1008477.
- 22 18 Lamarca AP, de Almeida LGP, Francisco R da S, *et al.* Genomic surveillance of SARS-
23 CoV-2 tracks early interstate transmission of P.1 lineage and diversification within P.2
24 clade in Brazil. *PLoS Negl Trop Dis* 2021; **15**: e0009835.
- 25 19 Chinazzi M, Davis JT, Ajelli M, *et al.* The effect of travel restrictions on the spread of the
26 2019 novel coronavirus (COVID-19) outbreak. *Science* 2020; **368**: 395–400.
- 27 20 Jorge DCP, Rodrigues MS, Silva MS, *et al.* Assessing the nationwide impact of COVID-
28 19 mitigation policies on the transmission rate of SARS-CoV-2 in Brazil. *Epidemics*
29 2021; **35**: 100465.
- 30 21 Fairchild G, Polgreen PM, Foster E, Rushton G, Segre AM. How many suffice? A
31 computational framework for sizing sentinel surveillance networks. *Int J Health Geogr*
32 2013; **12**: 56.
- 33 22 Pei S, Teng X, Lewis P, Shaman J. Optimizing respiratory virus surveillance networks
34 using uncertainty propagation. *Nat Commun* 2021; **12**: 222.

- 1 23 Naveca FG, Nascimento V, de Souza VC, *et al.* COVID-19 in Amazonas, Brazil, was
2 driven by the persistence of endemic lineages and P.1 emergence. *Nat Med* 2021; **27**:
3 1230–8.
- 4 24 Giovanetti M, Slavov SN, Fonseca V, *et al.* Genomic epidemiology of the SARS-CoV-2
5 epidemic in Brazil. *Nat Microbiol* 2022; **7**: 1490–500.
- 6 25 Gräf T, Bello G, Naveca FG, *et al.* Phylogenetic-based inference reveals distinct
7 transmission dynamics of SARS-CoV-2 lineages Gamma and P.2 in Brazil. *iScience*
8 2022; **25**: 104156.
- 9 26 Brazilian National Civil Aviation Agency (ANAC). <https://www.anac.gov.br> (accessed
10 Dec 20, 2023).
- 11 27 Instituto Brasileiro de Geografia e Estatística. Coordenac o de Geografia, Ligac es
12 Rodovi rias E Hidrovi rias, 2016. Rio de Janeiro: Rio de Janeiro: IBGE, Instituto
13 Brasileiro de Geografia e Estatística, 2017.
- 14 28 Brazilian National Transport Confederation.
15 <https://anuariodotransporte.cnt.org.br/2022/Inicial> (accessed Dec 20, 2023).
- 16 29 Ministry of Health, Brazil. Informatics Department Open Data SUS System.
17 <https://covid.saude.gov.br/> (accessed Dec 20, 2023).
- 18 30 Ford LR, Fulkerson DR. Maximal flow through a network. *Canad J Math* 1956; **8**: 399–
19 404.
- 20 31 Nakamura H, Managi S. Airport risk of importation and exportation of the COVID-19
21 pandemic. *Transport Policy* 2020; **96**: 40–7.

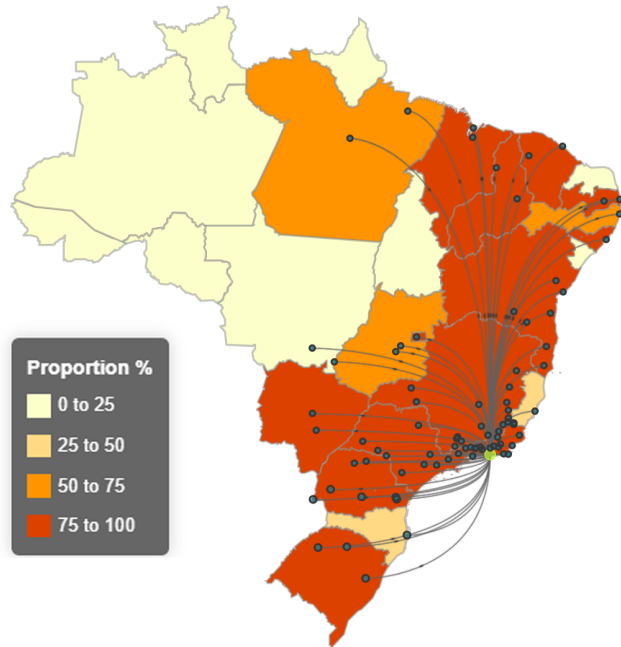




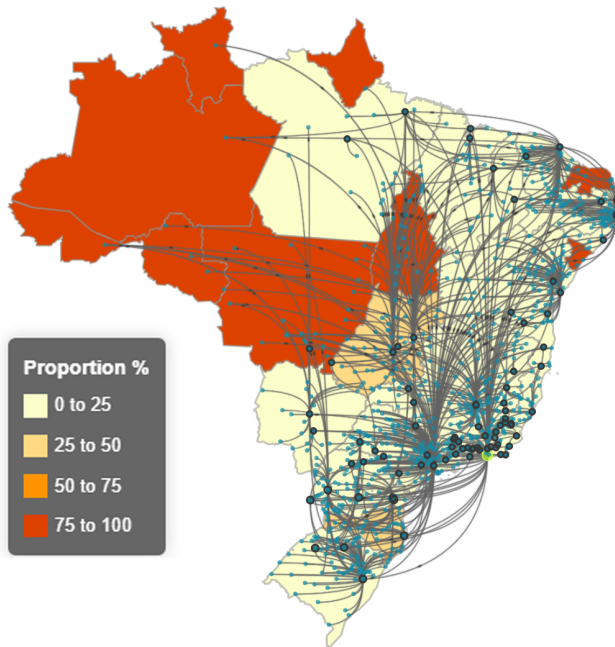


Rio de Janeiro city

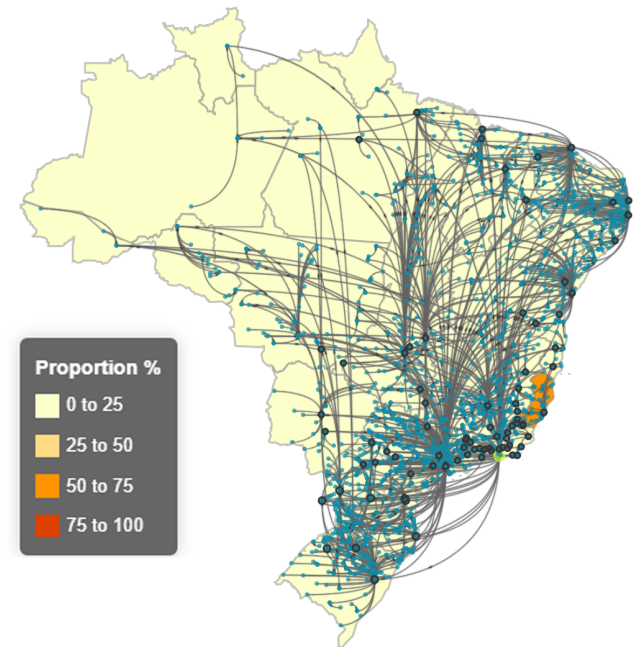
1st step cities



2nd step cities

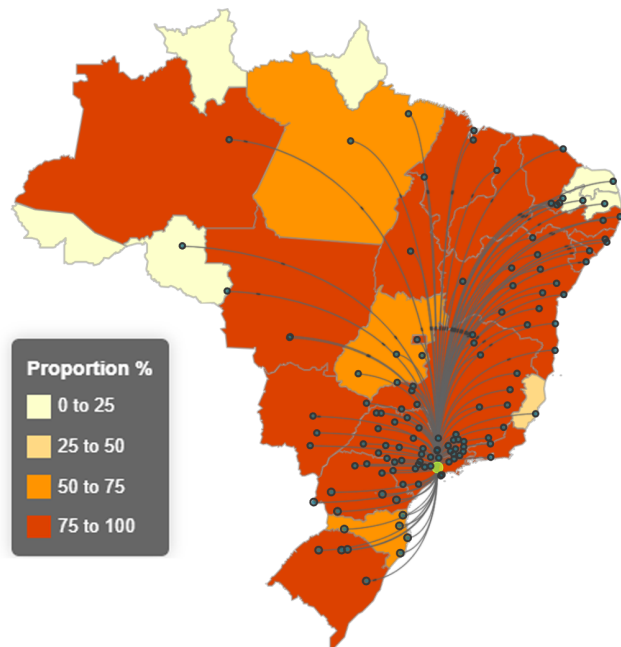


3rd step cities

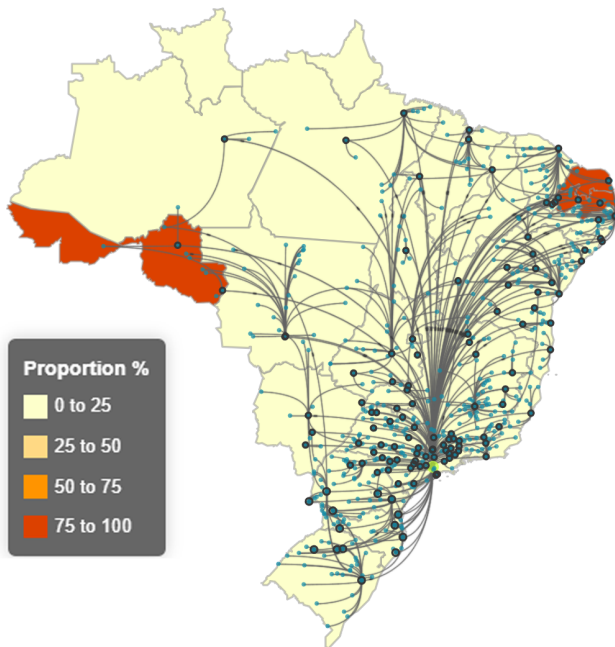


São Paulo city

1st step cities



2nd step cities



3rd step cities

