View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by El Servicio de Difusión de la Creación Intelectual

Uncovering the Spread of an Infectious Disease with Mobile Phone Data

Juan de Monasterio[1], Alejo Salles[2], Carolina Lang[3], Diego Weinberg[4], Martin Minnoni[5], Matias Travizano[5], and Carlos Sarraute[5]

[1] Mathematics Dept., Universidad de Buenos Aires
[2] Instituto de Cálculo, Universidad de Buenos Aires, and CONICET
[3] Computer Science Dept., Universidad de Buenos Aires
[4] Fundación Mundo Sano, Argentina
[5] Grandata Labs, Argentina

Abstract. We use mobile phone records for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. We show that geolocalized call records are rich in social and individual information, which can be used to infer whether an individual has lived in an endemic area. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country. The risk maps that we generate can be used by health campaign managers to target specific areas and allocate resources more effectively. Finally, we show the value of mobile phone records to predict long-term migrations, which play a crucial role in the spread of Chagas disease.

1 Introduction

Chagas disease is a tropical parasitic epidemic of global reach, spread mostly across 21 Latin American countries. The World Health Organization (WHO) estimates more than six million infected people worldwide [1]. Caused by the *Trypanosoma cruzi* parasite, its transmission occurs mostly in the American endemic regions via the *Triatoma infestans* insect family (also called "kissing bug", and known by many local names such as "vinchuca" in Argentina, Bolivia, Chile and Paraguay, and "chinche" in Central America). In recent years and due to globalization and migrations, the disease has become an issue in other continents [2], particularly in countries that receive Latin American immigrants such as Spain [3] and the United States [4], making it a global health problem.

A crucial characteristic of the infection is that it may last 10 to 30 years in an individual without being detected [5], which greatly complicates effective detection and treatment. About 30% of individuals with chronic Chagas disease will develop life-threatening cardiomyopathies or gastrointestinal disorders, whereas the remaining individuals will never develop symptoms. Long-term human mobility (particularly seasonal and permanent rural-urban migration) thus plays a key role in the spread of the epidemic [6]. Relevant routes of transmission also include blood transfusion, congenital contagion –with an estimated 14,000

newborns infected each year in the Americas [7]–, organ transplants, accidental ingestion of food contaminated by *Trypanosoma cruzi*, and even, in a minor scale, by laboratory accidents. The spatial dissemination of a congenitally transmitted disease sidesteps the available measures to control risk groups, and shows that individuals who have not been exposed to the disease vector should also be included in detection campaigns.

Mobile phone records contain information about the movements of large subsets of the population of a country, and make them very useful to understand the spreading dynamics of infectious diseases. They have been used to understand the diffusion of malaria in Kenya [8] and in Ivory Coast [9], including the refining of infection models [10]. The cited works on Ivory Coast were performed using the D4D (Data for Development) challenge datasets released in 2013. Tizzoni et al. [11] compare different mobility models using theoretical approaches, available census data and models based on CDRs interactions to infer movements. They found that the models based on CDRs and mobility census data are highly correlated, illustrating their use as mobility proxies.

Mobile phone data has also been used to predict the geographic spread and timing of Dengue epidemics [12]. This analysis was performed for the country of Pakistan, which is representative of many countries on the verge of countrywide endemic dengue transmission. Other works directly study CDRs to characterize human mobility and other sociodemographic information. A complete survey of mobile traffic analysis articles may be found in [13], which also reviews additional studies based on the Ivory Coast dataset mentioned above.

In this work, we discuss the use of mobile phone records –also known as Call Detail Records (CDRs)– for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. Key health expertise on the subject was provided by the *Mundo Sano* Foundation. We generate predictions of population movements between different regions, providing a proxy for the epidemic spread. Our objective is to show that geolocalized call records are rich in social and individual information, which can be used to determine whether an individual has lived in an epidemic area. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country. This is the first work that leverages mobile phone data to better understand the diffusion of the Chagas disease.

2 Chagas Disease in Argentina and Mexico

2.1 Key Facts and Endemic Zone in Argentina

For more than 50 years, vector control campaigns have been underway in Argentina as the main epidemic counter-measure. The *Gran Chaco*, situated in the northern part of the country, is hyperendemic for the disease [14]. A map of this ecoregion is shown in Figure 1. The ecoregion's low socio-demographic conditions further support the parasite's lifecycle, where domestic interactions between humans, triatomines and animals foster the appearance of new infection

cases, particularly among rural and poor areas. This region is considered as the endemic zone E_Z in the analysis described in Section 4 and Section 5.

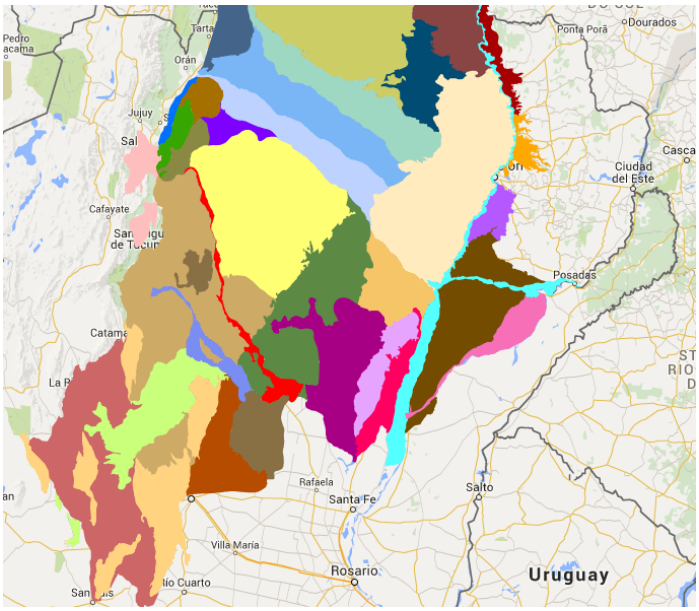


Fig. 1. The *Gran Chaco* ecoregion in South America.

The dynamic interaction of the triatomine infested areas and the human mobility patterns create a difficult scenario to track down individuals or spots with high prevalence of infected people or transmission risk. Available methods of surveying the state of the Chagas disease in Argentina nowadays are limited to individual screenings of individuals.

Recent national estimates indicate that there exist between 1.5 and 2 million individuals carrying the parasite, with more than seven million exposed. National health systems face many difficulties to effectively treat the disease. In Argentina, less than 1% of infected people are diagnosed and treated (the same statistic holds at the world level). Even though governmental programs have been ongoing for years now [15], data on the issue is scarce or hardly accessible. This presents a real obstacle to ongoing research and coordination efforts to tackle the disease in the region.

2.2 Key Facts and Endemic Zone in Mexico

In 2004, the joint work of *Instituto Nacional de Cardiología "Ignacio Chávez"* and *Instituto de Biología de la UNAM* resulted in a Chagas disease database for Mexico [16]. Reviewing positive serology in blood banks and human reported

Fig. 2. Endemic region E_Z for Mexico.

cases per state, an epidemic risk map description was produced to geographically situate the disease. Based on this data, we defined the Mexican epidemic area, selecting the states having the top 25% prevalence rates nationwide. The resulting risk region is shown in Figure 2. It covers most of the South region of the country and includes the states of Jalisco, Oaxaca, Veracruz, Guerrero, Morelos, Puebla, Hidalgo and Tabasco. This region is considered as the endemic zone E_Z for the Mexican case in the analysis described in Sections 4, 5 and 6.

The authors of [17] provide an extensive review of the research reports on Chagas disease in Mexico. The review is very critical, stating that there are no effective vector control programs in Mexico; and that the actual prevalence of the disease can only be estimated because no official reporting of cases is performed.

According to [18], there are a total of 18 endemic areas in Mexico, located in the southeast, and these areas include the states of Oaxaca, Jalisco, Yucatán, Chiapas, Veracruz, Puebla, Guerrero, Hidalgo, and Morelos, all of them with rural areas. Chiapas, Oaxaca, Puebla, Veracruz and Yucatán are among the most affected states (where the prevalence may exceed 10%), although cases have been reported in most areas of the country [16,18]. Despite the lack of official reports, an estimate of the number of *Trypanosoma cruzi* infections by state in the country indicates that the number of potentially affected people in Mexico is about 5.5 million [17]. Mexico, together with Bolivia, Colombia, and Central America, are among the countries most affected by this *neglected tropical disease* (NTD) [19]. The disease doesn't know about borders: Chagas and other neglected tropical diseases present in the north of Mexico remain highly endemic in the south of Texas as well [20].

In recent years there has been a focus on treating the disease with two available medications, benznidazole or nifurtimox. A study that explores the access to these two drugs in Mexico shows that less than 0.5% of those who are infected with the disease received treatment in Mexico in years [21].

People from endemic areas of Chagas disease tend to migrate to industrialized cities of the country, mainly Mexico City, in search of jobs. In accordance with this movement, a report showed that infected children under 5 year of age are frequently distributed in urban rather than in rural areas, indicating that the disease is becoming urbanized in Mexico [22]. Therefore, as in the Argentinian case, the study of long-term mobility is crucial to understand the spread of the Chagas disease in Mexico.

3 Mobile Phone Data Sources

Our data source is anonymized traffic information from two mobile operators, in Argentina and in Mexico. For our purposes, each record is represented as a tuple $\langle i, j, t, d, l \rangle$, where user i is the caller, user j is the callee, t is the date and time of the call, d is the direction of the call (incoming or outgoing, with respect to the mobile operator client), and l is the location of the tower that routed the communication. The dataset does not include personal information from the users, such as name or phone number. The users privacy is assured by differentiating users by their hashed ID, with encryption keys managed exclusively by the telephone company. Data was preprocessed excluding users whose monthly cellphone use either did not surpass a minimal number of calls μ or exceeded a maximal number M. This ensures we leave out outlying users such as call-centers or dead phones. In both datasets, we used $\mu = 5$ and $M = 400$.

We then aggregate the call records for a five month period into an edge list $(n_i, n_j, w_{i,j})$ where nodes n_i and n_j represent users i and j respectively and $w_{i,j}$ is a boolean value indicating whether these two users have communicated at least once within the five month period. This edge list will represent our mobile graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ where $\mathcal{N}$ denotes the set of nodes (users) and $\mathcal{E}$ the set of communication links. We note that only a subset $\mathcal{N}_C$ nodes in $\mathcal{N}$ are clients of the mobile operator, the remaining nodes $\mathcal{N} \setminus \mathcal{N}_C$ are users that communicated with users in $\mathcal{N}_C$ but themselves are not clients of the mobile operator. Since geolocation information is available only for users in $\mathcal{N}_C$, in the analysis we considered the graph $\mathcal{G}_C = \langle \mathcal{N}_C, \mathcal{E}_C \rangle$ of communications between clients of the operator.

Datasets Information. The Argentinian dataset contains CDRs collected over a period of 5 months, from November 2011 to March 2012. The raw data logs contain around 50 million calls per day. The Mexican data source is an anonymized dataset from a national mobile phone operator. Data is available for every call made within a period of 24 months from January 2014 to December 2015. The raw logs contain between 11 and 30 million calls per day for more than 8 million users that accessed the telecommunication company's (*telco*) network to place the call. This means that users from other companies are logged, as long as one of the users registering the call is a client of the operator. In practice, we only considered CDRs between users in $\mathcal{N}_C$ since geolocalization was only possible for this group.

4 Methodology for Risk Map Generation

In this section we describe the methodology used to generate risk maps for the Chagas disease in Argentina and in Mexico.

4.1 Home Detection

The first step of the process involves determining the area in which each user lives. Having the granularity of the geolocated data at the antenna level, we can match each user $u \in \mathcal{N}_C$ with its *home antenna* H_u. To do so, we assume H_u as the antenna in which user u spends most of the time during weekday nights. This, according to our categorization of types of days of the week, corresponds to Monday to Thursday nights, from 8pm to 6am of the following day. This was based on the assumption that on any given day, users will be located at home during night time [23,24]. Note that users for which the inferred home antenna is located in the endemic zone E_Z will be considered the set of *residents of* E_Z.

In the case of Argentina, the risk area is the *Gran Chaco* ecoregion, as described in Section 2.1; whereas in the case of Mexico, we used the region described in Section 2.2.

4.2 Detection of Vulnerable Users

Given the set of inhabitants of the risk area, we want to find those with a high communication with residents of the endemic zone E_Z. To do this, we get the list of calls for each user and then determine the set of neighbors in the social graph $\mathcal{G}_C$. For each resident of the endemic zone, we tag all his neighbors as potentially vulnerable. We also tag the calls to (from) a certain antenna from (to) residents of the endemic area E_Z as *vulnerable calls*.

The next step is to aggregate this data for every antenna. Given an antenna a, we will have:

- The total number of residents N_a (this is, the number of people for which that is their home antenna).
- The total number of residents which are vulnerable V_a.
- The total volume of outgoing calls C_a from every antenna.
- From the outgoing calls, we extracted every call that had a user whose home is in the endemic area E_Z as a receiver VC_a (*vulnerable calls*).

These four numbers $\langle N_a, V_a, C_a, VC_a \rangle$ are the indicators for each antenna in the studied country.

4.3 Heatmap Generation

With the collected data, we generated heatmaps to visualize the mentioned antenna indicators, overlapping these heatmaps with political maps of the region taken for study.

A circle is generated for each antenna, where the **area** depends on the population living in the antenna N_a; and the **color** is related to the fraction V_a/N_a of vulnerable users living there.

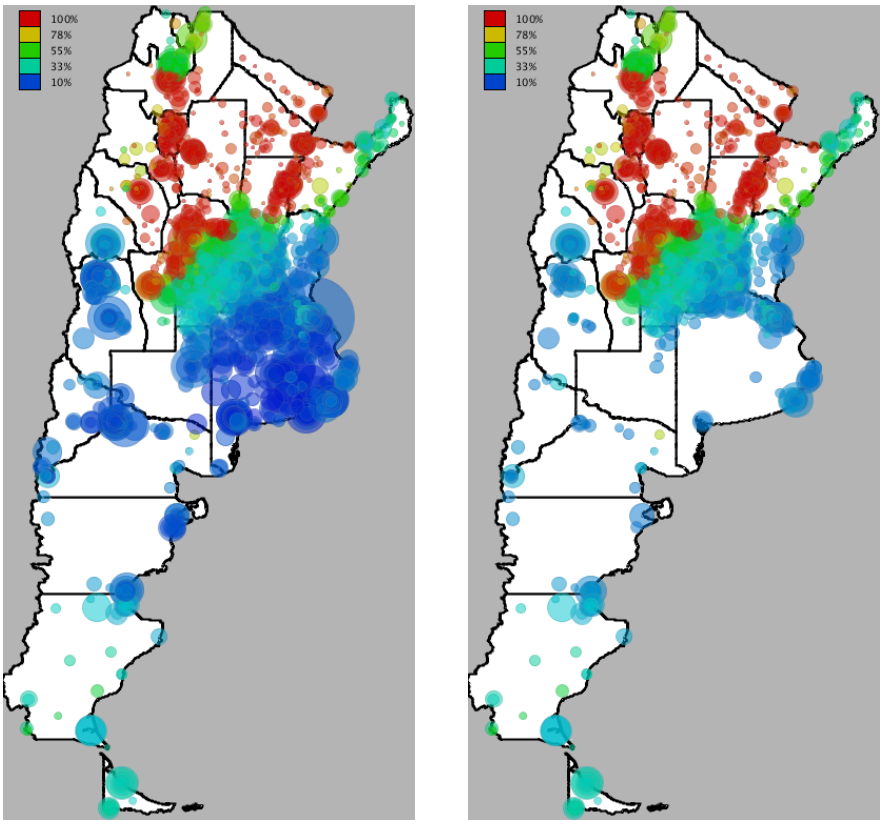
We used two filtering parameters to control which antennas are plotted.

- β: The antenna is plotted if its fraction of vulnerable users is higher than β.
- m_v: The antenna is plotted if its population is bigger than m_v.

These parameters were tuned differently for different regions. For example: an antenna whose vulnerable percentage would be considered low at the national level can be locally high when zooming in at a more regional level.

5 Results and Observations

5.1 Risk Maps for Argentina



(a) $\beta = 0.01$ (b) $\beta = 0.15$

Fig. 3. Risk map for Argentina, filtered according to β.

As a first visualization, maps were drawn using a provincial or national scale. Advised by *Mundo Sano* Foundation's experts, we then focused on areas of specific epidemic interest.

Figure 3 shows the risk maps for Argentina, generated with two values for the β filtering parameter, and fixing $m_v = 50$ inhabitants per antenna. After filtering with $\beta = 0.15$, we see that large portions of the country harbor potentially vulnerable individuals. Namely, Figure 3(b) shows antennas where more that 15% of the population has social ties with the endemic region E_Z.

Figure 4 shows a close-up for the Cordoba and Santa Fe provinces, where we can see a gradient from the regions closer to the endemic zone E_Z to the ones further away.

5.2 Zooming and Detection of Vulnerable Communities

As a result of inspecting the maps in Figure 3, we decided to focus visualizations in areas whose results were unexpected to the epidemiological experts. Focused areas included the provinces of Tierra del Fuego, Chubut, Santa Cruz and Buenos Aires, with special focus on the metropolitan area of Greater Buenos Aires whose heatmap is shown in Figure 5.

In some cases, antennas stood out for having a significantly higher link to the epidemic area than their adjacent antennas. Our objective here was to enhance the visualization in areas outside of Gran Chaco looking for possible host communities of migrants from the ecoregion. High risk antennas were separately listed and manually located in political maps. This information was made available to the *Mundo Sano* Foundation collaborators who used it as an aid for their campaign planning and as education for community health workers.

This analysis allowed us to specifically detect outlying communities in the focused regions. Some of these can be seen directly from the heatmap in Figure 5, where the towns of Avellaneda, San Isidro and Parque Patricios have been pinpointed.

5.3 Risk Maps for Mexico

With the data provided by the CDRs and the endemic region defined in Section 2.2, heatmaps were generated for Mexico using the methods described in Section 4. The first generated visualizations are depicted in Figure 6, which includes a map of the country of Mexico, and a zoom-in on the South region of the country. We used $m_v = 80$ inhabitants per antenna, and a high filtering value $\beta = 0.50$, which means that in all the antennas shown in Figure 6, more that 50% of inhabitants have a social tie with the endemic region E_Z. For space reasons, we don't provide here more specific visualizations and analysis of the regions of Mexico.

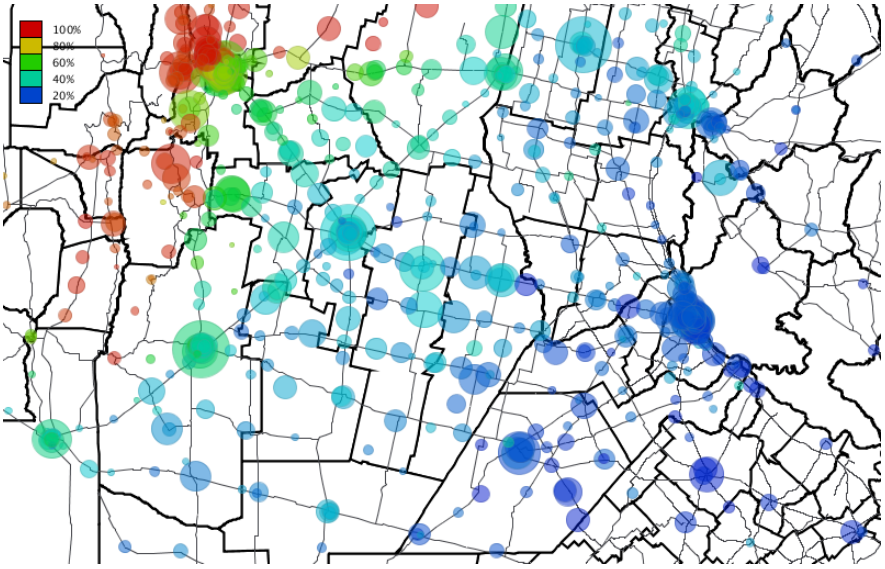


Fig. 4. Risk map for Cordoba and Santa Fe provinces, filtered according to $\beta = 0.15$.

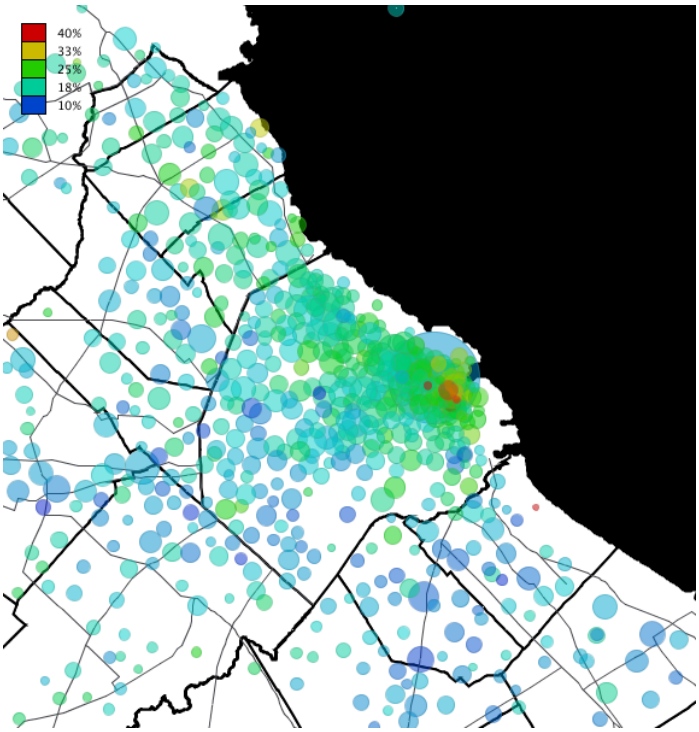
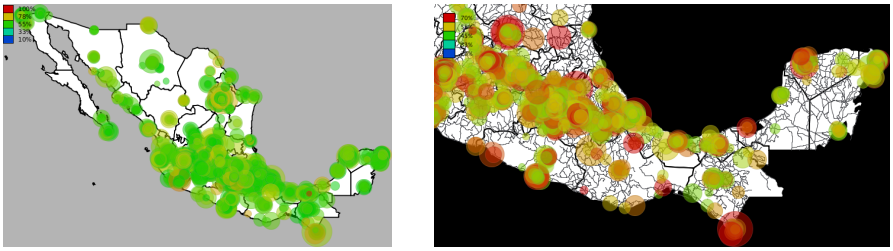


Fig. 5. Risk map for the metropolitan area of Buenos Aires, filtered with $\beta = 0.02$.



(a) National map, $\beta = 0.50$ (b) South region of Mexico, $\beta = 0.50$

Fig. 6. Risk maps for Mexico

6 Prediction of Long-term Migrations

In this section, we describe ongoing work on the prediction of long-term mobility. The CDR logs available for Argentina span a period of 5 months, whereas the Mexican dataset includes 24 months, from January 2014 to December 2015, making it more suitable for this study.

We divide the available data into two distinct periods: T_0, from January 2014 to July 2015, considered as the "past" in our experiment; and T_1, from August 2015 to December 2015, considered as the "present". Knowing which users live in the endemic region E_Z and how they communicate during period T_1, we want to infer whether they lived in E_Z in the past (period T_0). Our target variable Y is thus defined in the following way for every user u, where H_u is the user's home antenna:

$$Y_u = \begin{cases} 1 \text{ if } H_u \in E_Z \text{ during } T_0 \\ 0 \text{ in other cases.} \end{cases}$$

With this target variable, we tackle the prediction as a supervised classification problem.

6.1 Training Set

As explained, the training data belongs to period T_1, from August 2015 to December 2015; whereas the ground truth that we use to validate the predictions belongs to T_0, from January 2014 to August 2015. After preprocessing and cleaning the dataset, we obtained a training set with 1.6 million users.

Table 1 shows the percentage of antennas, the percentage of the population (according to INEGI census 2014), and the percentage of telco users per state, for the top 10 states.

The raw data logs contain between 11 million and 30 million calls per day and the volume of calls increases over the months, where most recent months have higher rates.

In this analysis we considered only postpaid users, i.e., users which have a monthly plan rate. This filtering was done because prepaid users have a higher churn rate, thus meaning that phone lines are not necessarily associated with

Table 1. Distribution of antennas, population and telco users by state.

State	Number of antennas	Population	Telco users
Distrito Federal	28.2%	8.5%	20.1%
Mexico	21.2%	13.9%	23.8%
Jalisco	10.7%	6.4%	8.3%
Nuevo Leon	9.6%	4.9%	2.9%
Guanajuato	6.1%	4.8%	5.9%
Puebla	5.8%	5.3%	4.3%
Veracruz	5.4%	6.8%	4.2%
Baja California	4.3%	2.8%	1.1%
Yucatan	4.1%	1.7%	2.9%
Sinaloa	4.1%	2.5%	0.4%

one single person during the two years of analysis, making them less suitable for the purpose of this study.

6.2 Model Features

The quality of the classification relies on the ability to characterize the users and their communication patterns. In general, the features constructed reflect calling and mobility patterns, segmented by different time periods during the week, and tagging whether the actions or subjects are 'endemic'. The training data runs from August 2015 to December 2015 (period T_1) and all CDRs are processed to extract features by user and by link.

Each week is divided into 3 time periods: (i) the period *weekday* is from Monday to Friday, on working hours (from 8hs to 20hs); (ii) *weeknight* is from Monday to Friday, between 20hs and 8hs of the following day; and (iii) *weekend* is Saturday and Sunday.

The model consists of the following features, which can be classified in 4 categories:

Used and home antennas. For each user $u \in \mathcal{N}_C$, we register the top ten most used antennas, during each month of the training period, together with the number of calls made through each antenna. We also register the most used antennas considering only calls made during the *weeknight* period, as defined above.

A user's home antenna is defined by the most used antenna during the *weeknight* period in all of T_0. From this, users were tagged as 'endemic' if their home antenna is in the endemic zone E_Z and 'exposed' if any of the ten antennas logged is in the risk area.

Mobility diameter. The user's logged antennas define a convex hull in space and the radius of the hull is taken to be as the mobility diameter. This length is

representative of the area of influence of that individual, a feature expected to be correlated with long-term migrations.

We register the mobility diameter of each user, as the diameter of the convex hull defined by his top 10 used antennas. Again, we generate two values, considering (i) all antennas and (ii) only the antennas used during the *weeknight*.

Graph data and communications. We look at the social graph $\mathcal{G}_C$ built from the CDRs, and the communications between nodes in $\mathcal{N}_C$.

For each edge $\langle n_i, n_j \rangle \in \mathcal{E}_C$, we dive into each of their interactions, segmenting call data with different criteria. For each month and each pair of users $\langle i, j \rangle$, we gather the tuple $\langle time_{ij}, calls_{ij}, dir, period \rangle$ where *time* is the sum of all calls (in seconds), *calls* is the number of calls exchanged, *dir* is a boolean variable indicating whether the calls were incoming or outgoing (from user i's point of view) and *period* corresponds to a segmentation of the week into the periods *weekday*, *weeknight*, and *weekend*.

Since the samples in our dataset are users, we have to aggregate all these variables, by grouping interactions at the user level. The combination of different variables amounts to a total of 130 features per user.

We also count each user's amount of neighbors in the communication graph and the total count of endemic neighbors, labeling each user i as *vulnerable* whenever he has any edge with another user j who lives in the endemic region E_Z.

Validation data. We perform an analysis similar to the home antenna detection previously described, but considering the time period T_0 (from January 2014 to July 2015), in order to determine the home antenna of users during T_0.

The number of people who maintain their home antenna between T_0 and T_1 is 1,012,416; whereas 580,425 users had a change in their home antenna. In terms of endemic condition, we observed that 1,551,560 users maintained their endemic condition between T_0 and T_1, whereas 41,281 had a change.

6.3 Supervised Classification

In this first iteration, we used the most common techniques found in the literature for this task: Support Vector Machines, Random Forest, Logistic Regression, and Multinomial Naive Bayes.

All algorithms were run on a 16-core Linux machine with 72GB of RAM. Processing and learning scripts were run on Python. Along the project, a variety of external packages were used for different purposes. For the classification routines, we used Scikit-learn [25] and Graphlab [26]. The data was split into 70% for training and 30% for testing.

Multinomial Bayes. The Multinomial Bayes classifier has a linear time complexity, and thus serves as a fast benchmark that we used to establish a baseline

classification performance. However, one shortcoming of this method is that it only allows for non-negative numerical features. This results in a reduced training and testing set.

The classifier has two hyperparameters: α, an additive smoothing parameter for which we defined values of $[0, 10^{-2}, 10^{-1}, 1]$; and the *fit prior* parameter which determines whether or not class prior probabilities are learned from the training set.

This setup gave 8 possible models and 24 fits on the 3-fold cross validated model training set where, on average, learning took 5s for one million samples. The F1-weighted metric was chosen to evaluate the best performing estimator. The best and worst scores achieved were 0.940 and 0.918 respectively.

SVM and Logistic Regression. Support Vector Machines (SVM) and Logistic Regression are two algorithms for classification which are standard in the Machine Learning literature and performed better than Multinomial Bayes in this case. At the same time, they are more complex and have higher time complexity, taking more computational resources to optimize.

Only the standard hyperparameters for these two models were tuned: $L2$-penalty regularization for Logistic Regression and kernel bandwidth for the Gaussian Kernel SVM. Both learning routines were executed in parallel and in each iteration 5% of the training set was sampled for cross validation. The best classifier was selected based on accuracy scores from this set.

The best model was a Logistic Regression Classifier with an $L2$-penalty value of 0.01. The following table shows the scores obtained by the selected model on the out-of-sample set.

Score	Value
F1 score	0.964537
Accuracy	0.980670
AUC	0.991593
Precision	0.970838
Recall	0.958316

High values across all scoring measures are achieved. These results can be explained by the fact that communication and mobility patterns are in essence highly correlated across time periods. In this case, correlation between the target variable Y and the models' features are expected: a user being endemic in T_1 is very correlated to being endemic in T_0, and the same holds with a user's interaction with vulnerable neighbors during T_1.

7 Conclusions and Future Work

7.1 Summary of Results

The heatmaps shown in Section 5 expose a "temperature" descent from the core regions outwards. The heat is concentrated in the ecoregion and gradually

descends as we move further away. This expected behavior could be explained by the fact that calls are in general of a local nature and limited to 3 or 4 main antennas used per user.

A more surprising fact is the finding of communities atypical to their neighboring region. They stand out for their strong communication ties with the studied region, showing significantly higher links of vulnerable communication. The detection of these antennas through the visualizations is of great value to health campaign managers. Tools that target specific areas help to prioritize resources and calls to action more effectively.

In Section 6, we tackled the problem of predicting long-term migrations. In particular, we showed that it is possible to use the mobile phone records of users during a bounded period (of 5 months) in order to predict whether they have lived in the endemic zone E_Z in a previous time frame (of 19 months). The very good results obtained demonstrate that CDRs are particularly well suited for this task.

To conclude, the results presented in this work show that it is possible to explore CDRs as a mean to tag human mobility. Combining social and geolocated information, the data at hand has been given an innovative use, different from its original billing purpose.

Epidemic counter-measures nowadays include setting national surveillance systems, vector-centered policy interventions and individual screenings of people. These measures require costly infrastructures to set up and be run. However, systems built on top of existing mobile networks would demand lower costs, taking advantage of the already available infrastructure. The potential value these results could add to health research is hereby exposed. Finally, the results stand as a proof of concept which can be extended to other countries or to diseases with similar characteristics.

7.2 Lines for Future Work

The mobility and social information extracted from CDRs analysis has been shown to be of practical use for Chagas disease research. Helping to make data driven decisions which in turn is key to support epidemiological policy interventions in the region. For the purpose of continuing this line of work, the following is a list of possible extensions being considered:

Results validation. Compare against actual serology or disease prevalence surveys. Data collected from fieldwork could be fed to the algorithm in order to supervise the learning.

Differentiating rural antennas from urban ones. This is important as rural areas have conditions which are more vulnerable to the disease expansion. *Trypanosoma cruzi* transmission is favored by rural housing materials and domestic animals contribute to complete the parasite's lifecycle.

Seasonal migration analysis. Experts from the *Mundo Sano* Foundation underlined that many seasonal migrations occur in the *Gran Chaco* region. The analysis of these movements can give information on which communities have a high influx of people from the endemic zone.

Search for epidemiological data at a finer grain. For instance, specific historical infection cases. Splitting the endemic region according to the infection rate in different areas, or considering particular infections.

Acknowledgements

We thank Marcelo Paganini, Marcelo Abril and Silvia Gold from Fundación Mundo Sano for their valuable input, useful discussions and support of the project. Special thanks to Adrián Paenza who provided the original idea and motivation of the project.

References

1. WHO. Chagas disease (American trypanosomiasis). *World Health Organization Fact sheets*, 2016.
2. Gabriel A Schmunis and Zaida E Yadon. Chagas disease: a Latin American health problem becoming a world health problem. *Acta tropica*, 115(1):14–21, 2010.
3. Miriam Navarro, Bárbara Navaza, Anne Guionnet, and Rogelio López-Vélez. Chagas disease in spain: need for further public health measures. *PLoS Negl Trop Dis*, 6(12):e1962, 2012.
4. Peter J Hotez, Eric Dumonteil, Miguel Betancourt Cravioto, Maria Elena Bottazzi, Roberto Tapia-Conyer, Sheba Meymandi, Unni Karunakara, Isabela Ribeiro, Rachel M Cohen, and Bernard Pecoul. An unfolding tragedy of chagas disease in north america. *PLoS Negl Trop Dis*, 7(10):e2300, 2013.
5. Anis Rassi and Joffre Marcondes de Rezende. American trypanosomiasis (Chagas disease). *Infectious disease clinics of North America*, 26(2):275–291, 2012.
6. Roberto Briceño-León. Chagas disease in the Americas: an ecohealth perspective. *Cadernos de Saúde Pública*, 25:S71–S82, 2009.
7. OPS. Estimacion Cuantitiva de la Enfermedad del Chagas en las Americas. *Organizacion Panamericana de la Salud/HDM/CD*, 425-06:1–6, 2006.
8. Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.
9. E Enns and J Amuasi. Human Mobility and Communication Patterns in Cote d'Ivoire: A Network Perspective for Malaria Control. *NetMob D4D Challenge*, pages 1–14, 2013.
10. R Chunara and EO Nsoesie. Large-scale Measurements of Network Topology and Disease Spread: A Pilot Evaluation Using Mobile Phone Data in Cote d'Ivoire. *NetMob D4D Challenge*, pages 1–18, 2013.
11. Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, 10(7):e1003716, 2014.
12. Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 112(38):11887–11892, 2015.

13. Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. *Mobile traffic analysis: a survey*. PhD thesis, Université de Lyon; INRIA Grenoble-Rhône-Alpes; INSA Lyon; CNR-IEIIT, 2015.
14. OPS. Mapa de Transmisión vectorial del Mal de Chagas. *Organizacion Panamericana de la Salud*, 2014.
15. Ministerio de Salud Argentina. Plan Nacional de Chagas, 2016. [Online, accessed 9-may-2016].
16. A Cruz-Reyes and José M Pickering-López. Chagas disease in Mexico: an analysis of geographical distribution during the past 76 years-A review. *Memorias do Instituto Oswaldo Cruz*, 101(4):345–354, 2006.
17. Alejandro Carabarin-Lima, María Cristina González-Vázquez, Olivia Rodríguez-Morales, Lidia Baylón-Pacheco, José Luis Rosales-Encina, Pedro Antonio Reyes-López, and Minerva Arce-Fonseca. Chagas disease (american trypanosomiasis) in Mexico: an update. *Acta tropica*, 127(2):126–135, 2013.
18. Eric Dumonteil. Update on Chagas' disease in Mexico. *Salud pública de México*, 41(4):322–327, 1999.
19. Peter J Hotez, Eric Dumonteil, Michael J Heffernan, and Maria E Bottazzi. Innovation for the 'bottom 100 million': eliminating neglected tropical diseases in the Americas. In *Hot Topics in Infection and Immunity in Children IX*, pages 1–12. Springer, 2013.
20. Peter J Hotez, Maria Elena Bottazzi, Eric Dumonteil, Jesus G Valenzuela, Shaden Kamhawi, Jaime Ortega, Samuel Ponce de Leon Rosales, Miguel Betancourt Cravioto, and Roberto Tapia-Conyer. Texas and Mexico: sharing a legacy of poverty and neglected tropical diseases. *PLoS Negl Trop Dis*, 6(3):e1497, 2012.
21. Jennifer M Manne, Callae S Snively, Janine M Ramsey, Marco Ocampo Salgado, Till Bärnighausen, and Michael R Reich. Barriers to treatment access for Chagas disease in Mexico. *PLoS Negl Trop Dis*, 7(10):e2488, 2013.
22. Carmen Guzmán-Bracho. Epidemiology of Chagas disease in Mexico: an update. *TRENDS in Parasitology*, 17(8):372–376, 2001.
23. Carlos Sarraute, Jorge Brea, Javier Burroni, Klaus Wehmut, Artur Ziviani, and Ignacio Alvarez-Hamelin. Social Events in a Time-Varying Mobile Phone Graph. In *Fourth International Conference on the Analysis of Mobile Phone Datasets (NetMob)*, 2015.
24. Balázs Cs Csáji, Arnaud Browet, VA Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 2012.
25. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
26. Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. Distributed graphlab: A framework for machine learning and data mining in the cloud. *Proc. VLDB Endow.*, 5(8):716–727, April 2012.