People, Quakes, and Communications: Inferences from Call Dynamics about a Seismic Event and its Influences on a Population

Ashish Kapoor[1] **Nathan Eagle**[2] **Eric Horvitz**[1]

[1]Microsoft Research
One Microsoft Way
Redmond WA 98052
{akapoor, horvitz}@microsoft.com

[2]The Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM87501
nathan@mit.edu

Abstract

We explore the prospect of inferring the epicenter and influences of seismic activity from changes in background phone communication activities logged at cell towers. In particular, we explore the perturbations in Rwandan call data invoked by an earthquake in February 2008 centered in the Lac Kivu region of the Democratic Republic of the Congo. Beyond the initial seismic event, we investigate the challenge of assessing the distribution of the persistence of needs over geographic regions, using the persistence of call anomalies after the earthquake as a proxy for lasting influences and the potential need for assistance. We also infer uncertainties in the inferences and consider the prospect of identifying the value of surveying the areas so that surveillance resources can be best triaged.

Introduction

Cellular phone networks have matured into well-developed and relatively widespread systems in developing countries with otherwise minimal infrastructure. While these pervasive cellular networks are continually generating call data records (CDR) for billing and maintenance purposes, we consider this infrastructure as an innervating sensor network that can be used for natural and human event detection. Methods for making inferences from anonymized CDR could provide guidance for detecting and reacting to natural disasters in remote geographic regions. Opportunities include making inferences about the nature and needs of people and populations facing acute challenges or at risk, about allocating scarce reconnaissance resources, and proactive decision making and actions to minimize hunger, thirst, and the spread of disease.

Beyond core inferences, we shall investigate the handling of the inevitable uncertainties in predictions. Varying densities of phones and cell towers and other factors may lead to varying levels of confidence in inferences from call data. Thus, inferential methods should include representations and machinery for capturing and propagating uncertainties about the inferences themselves. We show how we can coherently represent and propagate uncertainties and can use these uncertainties to prioritize the collection of new data, via computation of the value of making additional observations. Such computations of information value can be used to triage scarce resources available for reconnaissance. For example, the methods can be used to compose plans for surveying different regions for damages and needs, in order to achieve maximum relief within an available reconnaissance budget.

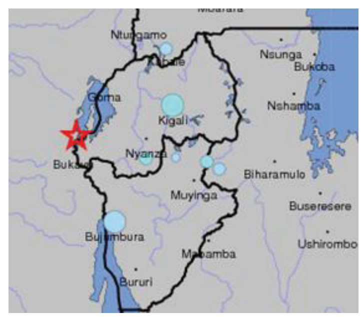

Figure 1. Location of the epicenter (star) of the February 2008 earthquake in the Lac Kivu region of the Democratic Republic of the Congo (courtesy US Geological Survey).

We evaluate such inferential opportunities in the context of the earthquake of February 3, 2008, centered in the Lac Kivu region of the Democratic Republic of the Congo. The epicenter and surrounding population centers as displayed in Figure 1. We shall examine the trends in call activity as logged at all cell towers within Rwanda before and after the earthquake and show how we can apply statistical modeling methods to: (1) detect when the earthquake occurred, (2) estimate the epicenter of the earthquake, (3) identify regions associated with persistence of anomalous activity, considering these regions and their respective population densities as a proxy for potential needs for assistance, and (4) quantify regional uncertainties to triage additional data collection efforts about the needs of a population.

We shall first present the general approach and overall framework for making such inferences, followed by the technical details of modeling and other computational considerations. Then, we focus on the case of the Lac Kivu earthquake and apply the methods to make inferences from the Rwandan CDR.

Related Work

Numerous research projects are leveraging the sensing capabilities of cellular phones and associated communication infrastructure as a resource of behavioral information. As examples, mobile phones have been used as sensors in determining social network structure [Eagle et al. 2009], performing activity recognition [Choudhary and Borriello 2008], and modeling human mobility [González et al. 2006].

Rather than building models of regular, recurring behavioral patterns, we pursue the detection and modeling of rare, disruptive events. With this approach, we consider background activity to learn patterns of normalcy, and then seek to detect and understand anomalies and their implications within small windows of time.

Approach

Assume that there are N cell towers and that for any i^{th} tower (where $i \in \{1..N\}$), we have a time series of observations: $\{a_i^1, a, \ldots a_i^t, \ldots a_i^T\}$ about communication activity on consecutive days $1..T$ logged by each cell tower. We also have access to the longitudes and the latitudes (x_i, y_i) of the geographical positions of these towers.

We shall consider observations jointly for all the cell towers and make inferences from changes in call volume that might have disrupted or influenced a population in some way. For certain kinds of events, we may be interested in inferring a central point of maximal intensity. For others, we may additionally wish to infer regions where maximal disruption to populations may have occurred. By making such inferences, we seek to build maps that would highlight areas where assistance or relief efforts might need to focus and where additional information is required before informed decisions can be made about resource allocation.

We shall make three assumptions in our analyses:

1. Cell tower traffic deviates statistically from the normal patterns and trends in case of an unusual event.

2. Areas that suffer larger disruptions experience deviations in call volume that persist for a longer period of time.

3. Disruptions are overall inversely proportional to the distance from the center(s) of a catastrophe.

Note that the first assumption talks about deviations, which can either be increase or decrease in call activity, consequently the approach based on assumption should be able to deal with events that induce both kinds of deviations. The second assumption is based on the observation that cell phones capture the pulse of human activity and discourse in a region. Following a large-scale event such as an earthquake, people may increase call traffic as they check in on safety, seek assistance, or coordinate in other ways. In other cases, a reduction of call traffic may occur given disruption to functionality of the phone system and large-scale loss of life. Regardless of the different mixes of these phenomena, we may often see anomalous call activity. The third assumption about centrality and diminishment with distance captures such disruptive phenomena as earthquakes that are often linked to a point of origin or epicenter.

Detecting Events. We shall first seek to build a system that can detect unusual events, such as disruptions caused by seismic events, by analyzing the background and dynamics of tower-level call volume. Let us assume that for every cell tower we have a Gaussian model that reflects regular activity. Formally for an uneventful day t,

$$p(a_i^t \mid Non - Event) \sim N(m_i, \sigma_i^2) \text{ where } i \in \{1..N\}$$

Here, m_i and σ_i^2 denote the mean activity and the variance of the i^{th} cell tower and can be estimated from historical data. Given this representation of normal activity, we can detect anomalous events by identifying deviations from the normal activity and trends in one or more cell towers. To detect unusual activity, we seek to identify how well the current observations fit the normal activity. Specifically, we shall employ the negative log likelihood as a scoring mechanism for detecting anomalies in call data:

$$Score^t = \sum_{i=1}^{N} \frac{(a_i^t - m_i)^2}{2\sigma_i^2} + log\sigma_i \quad (1)$$

A higher $Score^t$ reflects an increased likelihood of an anomalous event occurring on day t and this proposed measure can be used in a detection procedure.

Predicting Location of Event. Once we detect that an event of significance has taken place, we seek to identify the region at the center of the disruption or catastrophe from multiple cell towers. We shall rely on Assumption 3, which asserts that the call volume at towers that are closer to the vicinity of the central region of the disruption should have larger increases in activity. In particular, we assume that, in light of a significant event, the cell tower activity is influenced by the distance it is from the event center, (e_x, e_y). Formally, if $\theta \triangleq \{e_x, e_y, \alpha\}$ then we assert,

$$p_\theta(a_i^t \mid Event) \sim N(m_i + \frac{\alpha}{D_i^{(e_x,e_y)}}, \sigma_i^2) \quad (2)$$

Here, $D_i^{(e_x,e_y)} = \sqrt{(x_i - e_x)^2 + (y_i - e_y)^2}$ denotes the distance of the i^{th} cell tower from the center (e_x, e_y) and α is an unknown scaling parameter. Given this model and the observations on the day of the event, we invoke the principle of maximum likelihood to estimate the unknown center of action, (e_x, e_y), and the scaling parameter. In

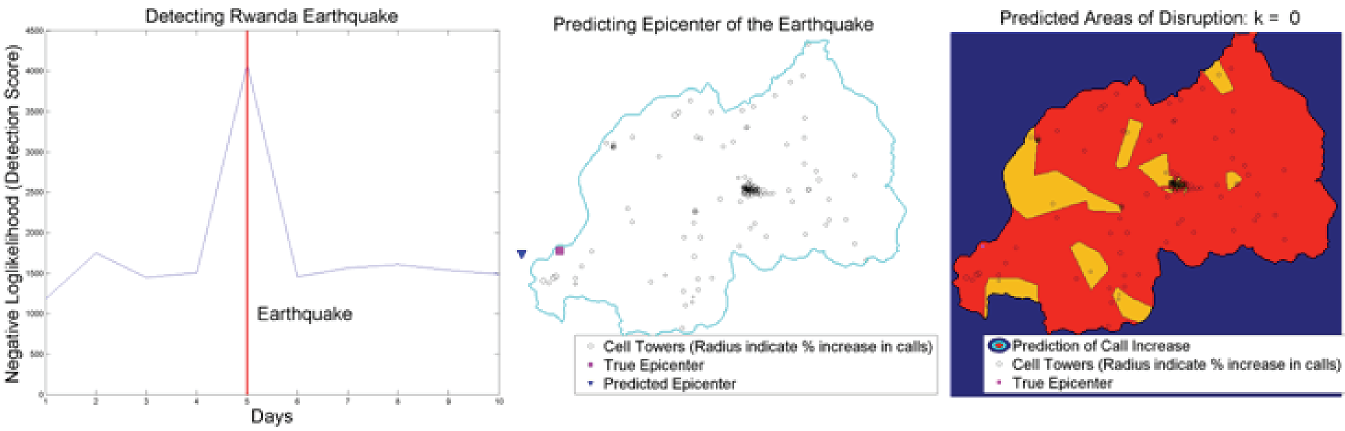

Figure 2. Left: Detection of event from cell tower data. Middle: Scatter plot with predicted epicenter, true epicenter, and tower activity. Right: Predicted regions where call traffic exceeds 1-σ from the baseline (warmer/darker shades) for the day of the event.

particular, we search for the best θ^* to estimate the epicenter by maximizing the log likelihood,

$$\theta^* = \text{argmax}_\theta \sum_{i=1}^{N} \log p_\theta(a_i^t \mid Event)$$

We can solve for θ^* via search, using gradient-descent optimization to determine the parameters (e_x, e_y) and α, thus inferring a central location of a disruptive event.

Predicting Opportunities for Assistance. Beyond identifying regions where there are acute changes in call activity in response to a disruptive event, we wish to make inferences about regions that have likely suffered more damage and thus are higher priority areas of attention for the provision of assistance. Beyond triaging attention, we are also interested in opportunities to make direct inferences about the nature and geographical distribution of ideal sets of proactive actions that might be taken for such goals as maximizing the survival of people who have been injured or are trapped, coalescing transportation resources and expertise for medical care, and creating, readying, and perhaps even implementing contingency plans for transporting medications, food, and water. The latter can be important with minimizing or ameliorating the spread of such diseases as cholera, which may follow natural disasters with some delay. Cholera has a 5% mortality rate in Africa and the primary treatment is the provision of sufficient water to patients. We are interested in opportunities to construct predictive models that can identify regions at risk for a jump in cholera incidence following a disastrous event. Proactive measures guided by predictive models, such as preparing to ensure that water and related medical assistance is available for transport to such regions, could reduce morbidity and mortality.

Per Assumption 2, we shall consider a significant and persistent deviation from the baseline in call volume, as a signal of disruption. Our strategy is to build a model that can accurately predict if a significant deviation in a tower's call volume would persist. Given that many people may communicate by phone to simply check in with family and friends about the acute influences of an earthquake, we wish to consider the region-specific *persistence* of anomalous activity over time as a proxy for significant disruption and as an indication of opportunities for assistance in those regions.

We seek to identify whether a seismic event will lead to changes in call activity that persist days after the event as a sign of persistent needs. We explore a predictive model that considers cell tower coordinates, in conjunction with the prior activity and population around towers, to predict whether a significant deviation of call activity from baseline will persist.

Let us assume that a disruptive event occurred on day t'. We are interested in predicting whether a significant increase in activity at an i^{th} cell tower would be observed at k days following the event. To this end, we train a classifier w, that predicts anomalous cell traffic at k days, given activity at the cell towers.

Let us consider call activity a_i to be a significant deviation if the call traffic differs by more than one-sigma (σ_i) from the mean m_i at baseline. We shall consider three observations for each cell tower. Formally, $o_i = \left[[\tilde{a}_i;\ \pi_i; D_i^{(e_x,e_y)}; x_i; y_i\right]^T$ denotes the observation vector corresponding to the i^{th} tower and the features represent deviation in activity on the day of earthquake $(\tilde{a}_i \triangleq \frac{a_i^t - m_i}{\sigma_i})$, the population density (π_i) around the tower, its distance $(D_i^{(e_x,e_y)})$ from the center of the event, and its coordinates (x_i, y_i).

We include the population density as an evidential feature so as to capture the prospect that damage and disruption in a region that is buffeted by seismic forces is a function of the density of people living in regions. With increasing density of a population comes increasing densities of dwellings, and greater numbers of people influenced by the breach of structural integrity of buildings

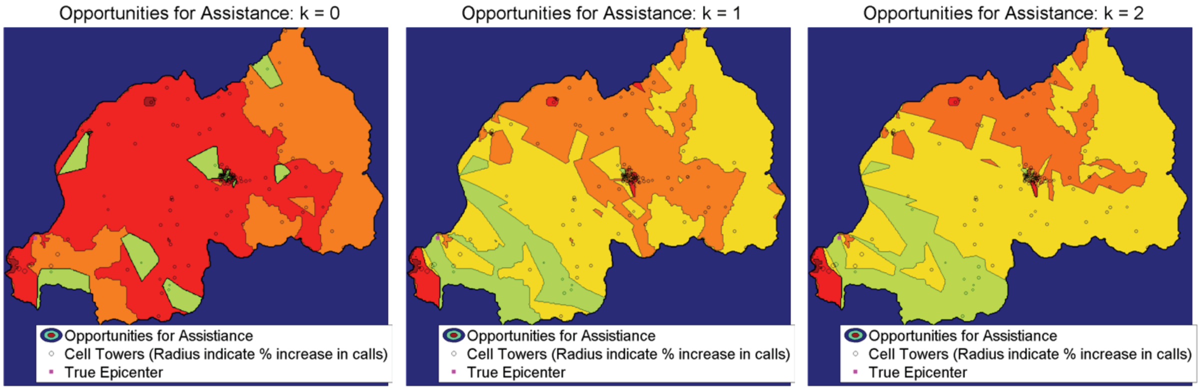

Figure 3. Inferences of opportunities for assistance. Maps display predictions about regions associated with increased opportunities for assistance, using as a proxy for disruption the extension of anomalous call activity to *k* days following the earthquake, weighted by population density. Warmer colors (darker shades) correspond to regions with increased opportunities for assistance.

and related infrastructure. Also, the costs of diminishment of flows of food and water may rise rapidly in densely populated regions. Greater densities of population can also raise the risk of transmission of disease.

Given these features and the data collected from the towers, we can infer a linear classifier w using algorithms such as Support-Vector Machines, logistic regression, and Gaussian Process classification (GPC) [Rasmussen and Williams 2006]. We shall review the construction of a classifier with GPC as it provides both predictions and estimates of uncertainties about those predictions. As we shall see, this classifier can be used to compute the value of information, which we employ later to compute the value of surveying regions. Formally, building the classifier results in a most-likely classifier, represented as w, and the variance around it Σ_{w}. For any test point, o_{test}, the predictive probability of persistence can be written as:

$$p_{test}^{persist} = \Phi\left(\frac{w^T o_{test}}{\sqrt{1 + o_{test}^T \Sigma_w o_{test}}}\right)$$

Here $\Phi(\cdot)$ denotes cumulative distribution function (cdf) of a normal distribution. This model thus can be used to persistence of a significant deviation at any hypothetical cell tower located at a coordinate (*x,y*). Specifically, given the parameters (w, Σ_{w}) and the test location (*x,y*), we can predict the persistence of deviation in the hypothetical cell tower at a particular location. In use, we compute the distance from epicenter and obtain the population density $\pi_{(x,y)}$. Then, we approximate $\tilde{a}_{(x,y)}^{t'}$, the deviation of activity of hypothetical towers on the day of the earthquake. We use a nearest-neighbor approximation, where we identify an existing cell tower j that is nearest to the location *(x,y)* and assume that $\tilde{a}_{(x,y)}^{t'} = \tilde{a}_j^{t'}$.

We compute an *assistance-opportunity score* for characterizing opportunities for assistance and identifying regions that might most benefit by relief efforts. Under the assumption that areas with high population density require more relief effort per unit of region we define the assistance score as the product of predicted relative increase in persistent call traffic multiplied by the population density:

$$AssistScore_{(x,y)} = p_{(x,y)}^{persist} \cdot \pi_{(x,y)}$$

Determining Value of Survey. Cell towers are most densely packed near big cities, capitals, and overall more developed parts of countries. Consequently, we can expect to have more confidence in predictions about opportunities for assistance around the areas with higher cell tower density, and have less confidence about inferences based on fewer cell towers. Such uncertainty can be reduced with the pursuit of additional information following an earthquake. However, as surveillance resources are scarce and costly, we pursue a formal model for triaging scarce reconnaissance resources under a limited budget. We take a decision-theoretic perspective to compute the *expected value of surveying* a region by considering expected benefit and costs of gathering information. In particular, we seek to select a set of locations S^* from the set of non-instrumented locations U that provide maximum gain per unit cost[1]:

$$S^* = argmax_{S \subseteq U} \frac{Gain(S)}{Cost(S)}$$

Given inferences that provide uncertainties in predictions, we can compute the *expected value of information* [Howard, 1967; Horvitz, Breese, and Henrion, 1989]. We shall define $Gain(S)$ as reduction in uncertainty at non-instrumented locations. Formally, we use A to denote the set of locations that we have information about and U as the set we have not surveyed, respectively. We write the selection criterion as:

[1] This criterion can also be represented as $Gain\ S - Co(S)$; using gain per unit cost enables allows gain and the cost to be in different currencies.

$$S^* = argmax_{S \subseteq U} \frac{H(U-S|A) - H(U-S|A \cup S)}{Cost(S)}$$

where $H(\cdot)$ denotes entropy. This formulation attempts to find the set S that provides maximum information about the rest of the sites (denoted as: $U-S$) under minimum costs. It is known that determining S^* is computationally intractable for a large set U, however, a greedy solution to this problem results in a close approximation to the ideal solution in settings where a sub-modularity property holds [Krause et al. 2008]. We note that the above mentioned criteria attempts to optimize a gain in terms of reduction in uncertainty, without taking into account either the amount of disruption or the expected gain in terms of human lives that could be saved. Beyond optimizing the reduction in uncertainty, we can consider a gain, which we call expected value of survey, by multiplying the information theoretic savings ($\Delta H(\cdot)$) with the population density and the expected disruptions ($p_s^{persist}$). Formally, the greedy selection procedure selects the location *s* to survey that maximize the following:

$$ValueOfSurvey_s = p_s^{persist} \cdot \pi_s \cdot \frac{\Delta H(s)}{Cost(s)}$$

The other detail we need is an estimation of the uncertainty about inferences. For linear Gaussian Process models [Rasmussen and Williams 2006], we can show that the information theoretic gain can be written as [Krause et al. 2008]:

$$\Delta H(s) = \log \left(\frac{K_{ss} - K_{sA} K_{AA}^{-1} K_{As}}{K_{ss} - K_{s\bar{A}} K_{\bar{A}\bar{A}}^{-1} K_{\bar{A}s}} \right)$$

Here, $K = [k_{ij}]$, is a kernel matrix where $k_{ij} = o_i^T o_j$ are the linear projections.

Finally, we can approximate the cost of surveying a location as a function of the distance from a major city. However, we emphasize that cost can be modeled using various factors such as geography, financial considerations, time to respond, and other relevant variables. We can sequentially select sites to survey in a greedy manner until the budget is exhausted. Thus, given the location of cell towers and logged call activity, we can use the above methodology to determine the areas that should be probed under a budget in order to best triage relief efforts.

Results

We now test the proposed framework in the context of Rwandan CDR. In particular, this data is aggregated to the tower-level, consisting of daily, directed communication volume for each cell tower in the country over a period of 3 years. These include data during the week including February 3, 2008, when a 5.9 magnitude earthquake was observed with an epicenter located by the USGS at 2.318 S and 28.945 E.

We first start by building baseline models from historic data recorded during a normal time-period. In particular we look at a continuous period of ten days and for each i^{th} cell tower record the mean m_i and the variance σ_i^2. This constitutes a baseline model and we use this model in performing the computations as described earlier.

Detecting the Earthquake. We use the event detection score as described in Equation 1 to determine deviation from normal activity. Figure 2 (left) shows the scores for 10 consecutive days around Feb 3, 2008. We can see that the score spikes at the correct day when the earthquake occurred demonstrating that such a scoring scheme can be used to detect seismic events.

Predicting Seismic Epicenter. We next pursue the challenge of predicting the location of the epicenter from the cell tower activity. We use the model described in Equation 2, and use the communication data to infer the epicenter. In particular we maximize the likelihood of the model for this challenge. Figure 2 (middle) shows the result of this experiment. The cell towers are depicted as black circles with radii indicating call activity handled by cell towers. We plot both the epicenter identified by USGS (magenta square) and the predicted epicenter. The predicted epicenter (-2.34, 28.71) is in close proximity to the USGS epicenter (-2.32, 28.94), highlighting the promise of using call activity and the existing communication infrastructure as a large-scale seismic sensing system.

Inferring Opportunities for Assistance. We also seek to employ geospatial methods to model persistence of deviations in cell tower traffic. As described earlier, modeling this persistence may help to identify regions where relief efforts are most needed. For experimental purposes, we learn the geospatial model parameters for *k = 1,..,5* days. More specifically, to explore capabilities of the model we use the tower data to build predictive models for each of five days following an earthquake.

We perform leave-one-out analysis in order to verify the performance of the model. In particular, for every cell tower in the training set we build a leave-one-out model using the rest of the training data and then use the model to predict the classification label the tower that has been left out (label = +1 means whether a significant effect persists or not). Table 1 shows recognition results using leave-one-out and compares it with a baseline approach of using the observed activity on the day of the earthquake as predictions for persistence. We also mention the marginal rates (maximum recognition obtained when the classifier predicts same label for all the towers). We can see that the predicted model is superior to the baseline and provides predictions that are significantly better for *k =3, 4* and *5.*

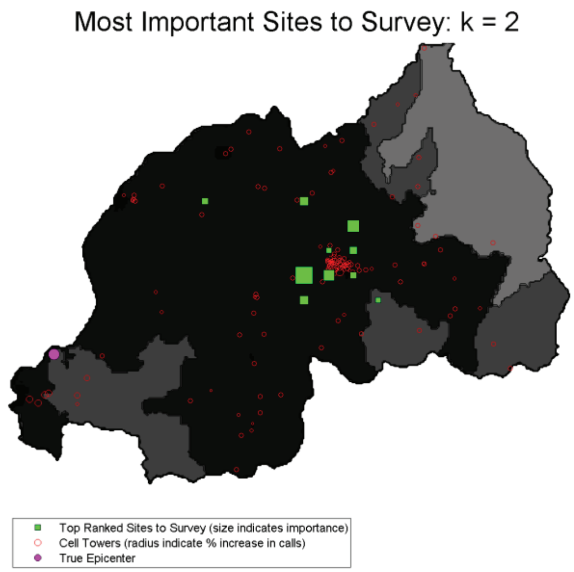

Figure 4. Top ten sites to survey (green squares), with rank indicated by the size of square. Magenta dot is true epicenter.

.

Table 1. Performance of geospatial-temporal model and baseline use of the previous day's observations.

k	Accuracy (Predictive Model)	Accuracy (Baseline)	Marginal
1	**0.63**	0.54	0.57
2	**0.60**	0.50	0.50
3	**0.78**	062	0.71
4	**0.74**	0.62	0.62
5	**0.65**	0.45	0.57

We apply the learned model to predict observations for any location (x, y), consequently recovering an estimate of disruption. Figure 2 (right) displays a map showing these predictions. Regions near the epicenter show higher disruption. However, disruptiveness is not smoothly distributed. As the model encapsulates the population density and call activity, we obtain a richer view of regions of disruption, per the definitions we have formulated.

We use these predictions about anomalous call volumes to compute the $AssistScore$ at all locations. Figure 3 highlights the regions that we infer would most benefit from relief efforts, based a definition of disruption as call traffic anomalies at k days following the earthquake. As we shift the definition of "disruption" as the extension of call traffic anomalies to increasingly longer durations, the inferred regions of increased opportunities for assistance shift away from epicenter, toward other regions of the country.

Inferring Regions to Survey. Next, we explore the potential value of predictive modeling in computing the value of survey. We employ the greedy information-value procedure to select the top ten sites that should be surveyed in order to make the relief efforts effective. In this experiment, we assume that the cost of surveying a site is directly proportional to its distance from Kigali.

Figure 4 shows the map of the country with the top ten sites to survey. The figure also shows the existing cell towers. Further, the gray levels of different areas correspond to population density. Again, if we had used predictions to model call activity only on the day of the earthquake, the majority of the predicted regions to survey would be near the earthquake center. However, modeling the extension of disruptions to additional days, and considering anomalous call traffic at later days as proxies for disruption, leads to recommendations to survey much wider areas, especially for regions with high population density but fewer cell towers.

Conclusion and Future Work

We presented methods for using the cellular phone infrastructure to detect seismic events and their influences on a population. We applied the methods to tower-level CDR from Rwanda and demonstrated our ability to detect the 2008 Lac Kivu earthquake and estimate its epicenter. We reviewed approaches to inferring regions that require relief efforts and for guiding surveys. The results highlight the promise of performing predictive analyses with existing telecommunications infrastructure. Future research directions include running sensitivity analyses over ranges of parameters and assumptions to explore the robustness of the results, the use of richer models that consider such information as geographic terrain and more detailed measures of seismic activity, and methods for guiding proactive planning, aimed at mitigating such downstream phenomena as the cutting of food supply lines and the outbreak of disease.

References

Choudhury, T., Borriello, G. The Mobile Sensing Platform: An Embedded System for Activity Recognition. IEEE Pervasive Magazine (2008).

Eagle, N., Pentland, A. and Lazer, D. Inferring Social Network Structure using Mobile Phone Data. Proc. of the NAS (2009).

González, M. C., Hidalgo, C. A., Barabási, A.-L. Understanding individual human mobility patterns. Nature 453, 779-782 (2008).

Horvitz, E., Breese, J.S., Henrion, M. Decision Theory in Expert Systems and Artificial Intelligence. Intl. J. Approx. Reasoning (1988).

Howard, R.A., Information value theory. IEEE Trans. on Systems Science and Cybernetics (1966).

Krause, A, Singh, A. and Guestrin, C. Near-optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. JMLR (2008).

Rasmussen. C. E. and Williams, C. K. I. Gaussian Processes for Machine Learning. MIT Press (2006).