Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 00 (2015) 000–000

Procedia Engineering

www.elsevier.com/locate/procedia

Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015

Comment [S1]: Elsevier to update with vo and page numbers.

HackEbola with Data: On the hackathon format for timely data analysis

HackEbola with Data[a,]*

[a]*Full author list& affiliations in Section 6*

Abstract

Many humanitarian organizations have been organizing data relating to the current Ebola epidemic in West Africa. These efforts include not only collecting statistics on Ebola cases and deaths, but also consolidating information about the hardest hit countries—Liberia, Guinea, and Sierra Leone—at a sub-regional level. During HackEbola with Data, participants were tasked to use this data to address the following question: *What factors are affecting the regional and temporal evolution of the Ebola epidemic in West Africa?* We report on both the specific findings from HackEbola with Data and lessons learned on using such a format to address complex data analysis needs in future crises.

Peer-review under responsibility of the Organizing Committee of HumTech2015.

Keywords: Ebola; Hackathons; Data Analysis

1. Motivation and objective

Many humanitarian organizations have been collecting, consolidating, and making public data relevant to the Ebola epidemic in West Africa. Data aggregators, such as the Humanitarian Data Exchange and the Open Humanitarian Data Repository, have not only made available sub-regional time-series of Ebola cases and counts from various organizations, but they are also a source for a variety of other sub-regional characteristics, such as socioeconomic indicators, food prices, and infrastructure types. These additional sources create the opportunity to

* Corresponding author.
E-mail address: finale@seas.harvard.edu

Peer-review under responsibility of the Organizing Committee of HumTech2015.

go beyond standard models that consider only case and fatality rates; they allow us to explore what factors may be affecting the outbreak. While understanding these factors may result in actionable insights, analyzing such complex data in a timely manner is a challenging task.

Hackathons are a popular event format, especially in computer science and engineering fields. A hackathon is a short event, usually taking place over a day or a weekend, usually centered around an open-ended goal. These goals can be very broad, such as creating interesting smart-phone applications, or more focused, such as building classification models for biological phenotypes based on high throughput or genomics data. While there may be prizes involved, participants are often drawn to hackathons as a way to do some technical work in a positive, high-energy atmosphere. In this paper, we describe the main findings from HackEbola (with Data!), a weekend-long event to assist in the Ebola data-analysis effort through the following question:

What factors may be affecting the regional and temporal evolution of the Ebola epidemic in West Africa?

We also discuss the use of a hackathon format to perform challenging data-analyses in a timely manner. HackEbola (with Data!) created an opportunity for talented data scientists to set aside a focused period of time to address this important data-analysis question. It also allowed for many analysis approaches to be tested quickly. Harvard University and the University of Massachusetts at Amherst both hosted local events, and many others participated remotely. qDatum.io provided the data management. Over the hackathon weekend, there were 2395 data downloads from 189 unique users. Over 300 people registered for the event, and 16 teams submitted analyses.

The objective of this paper is two-fold. First, we describe and discuss the use of a hackathon format for addressing pressing data analysis needs (Section 2). Next, we detail our technical outcomes, as well as the data sources, analysis approaches, and limitations (Section 3). We conclude in Section 4.

2. The hackathon format: Rationale and execution

Hackathons are a popular format, and there are many guides that describe how to run one. Here we focus on hackathons as a means to connect humanitarian organizations with data analysis needs to a community of highly-qualified data experts that may not otherwise know of them.

2.1. Background: Methods for connecting volunteer analysts to important needs

There exist many ways to find volunteers to assist with data analyses. Organizations such as Statistics Without Borders, connect specific statisticians with specific clients. The vetting process ensures a strong connection between a qualified statistician and a legitimate client. However, the client only has a few people assisting with the data analysis; this format may not be ideal for open-ended questions where one may wish to leverage the creativity of a larger group of data experts. Other organizations such as Kaggle or DrivenData issue public challenges or competitions. These allow a much broader community of people to participate. Such challenges are designed so that the merit of any solution can easily be computed and compared to other solutions; participants are often motivated by prizes and the competitive nature of the challenge. However, requiring that solution quality to be easily quantified—and ensuring that mechanisms (such as data obfuscation) are in place to prevent cheating—can limit the types of questions that are suitable for such a format. Both the consulting approach (e.g., SWB) and the competitive approach also typically occur over a longer period of time (weeks to months). This means that participants have more time to think about ideas that might be relevant to the client or challenge, and it also means that their energy devoted to the project is more diffuse.

In contrast, a hackathon involves a group of people working in a shared space toward some technical objective in a short period of time (usually a day or a weekend). Working in a shared space generates a shared energy, and emphasis is placed on creating a positive, high-energy working atmosphere. By limiting the duration of the event, participants can usually commit to an intense period of focused work toward the technical objective. The specificity of this objective varies depending on the event: some hackathons might simply task participants to create an interesting smart-phone application, while others may task participants with a much more focused objective, such as building tools to analyze a specific set of high-throughput genomics data. While prizes may be awarded, at its core a hackathon is a time for participants to get together and have fun "hacking." Participants often take part for the experience itself, rather than a prize or other reward; they are popular among computer scientists, engineers, and other non-domain experts who might just want to spend some time thinking and working on an interesting problem. Finally, hackathons differ from other short duration events, such as crowd-sourced data-labeling or processing efforts, in that they are typically more open-ended and expect some kind of creativity from the participants.

2.2. Rationale for the hackathon format for HackEbola with Data

There are many efforts to model the temporal evolution of the Ebola outbreak, and our goal was to find a way to contribute, rather than duplicate, the epidemiological analyses already underway. We chose a hackathon format in part because it would attract a broad variety of thinkers—data experts from all fields—to think creatively about new analyses to try. The hackathon format provided an excellent way for testing these varied modelling assumptions quickly. For example, teams varied in how they processed and summarized the spatially-explicit time series of cases and deaths; one could take the results that were consistent across these assumptions for closer scrutiny.

HackEbola (with Data!) was focused on producing insights; our non-competitive atmosphere encouraged participants to post cleaned up versions of the data and preliminary results for others to use and verify. Working in a shared space meant that participants could immediately disseminate their issues and insights. The non-competitive nature of the event also allowed participants to focus on more general questions like *how* and *what* rather than only on prediction, which are crucial in data exploration. Discussion of information and ideas between teams throughout the event also allowed greater efficiency by avoiding redundancy in project types and sharing solutions to common sub-problems (e.g., the need to mine basic demography parameter values).

Finally, HackEbola (with Data!) engaged data scientists at all levels of experience and raised awareness of the needs and challenges of analyzing data associated with a humanitarian crises. Several teams reported both that they had learned valuable lessons from HackEbola (with Data!) and that the experience had made them more likely to participate in similar events in the future.

2.3. Event execution

Prior to HackEbola (with Data!). Our co-organizers and data management partner qDatum.io created feeds of public data from Humanitarian Data Exchange and the Open Humanitarian Data Repository. They also ensured that the sub-regional administrative codes were consistent across the various data sets and mapped any positions that were included as latitude-longitude values to administrative regions. This data management and pre-processing was critical to the success of the event, as the data could then be directly imported into a variety of analyses tools. The consistent geographic identifiers allowed participants without any geography background (the majority of the participants) to easily link across the different data tables. We also provided advance background reading and video media on Ebola–valuable for data analysts without much epidemiology experience–as well as early access to the data.

Hacking. HackEbola (with Data!) began with a Friday evening kick-off dinner, during which participants were introduced to the data sets and the hackathon objectives (detailed in Section 3). Colby Wilkason, a crisis management expert from the Red Cross, gave a presentation on their needs. Participants formed teams through ChallengePost and began work. Saturday was also devoted largely to work, with meals and snacks being provided by the event sponsors. While there was a formal update session during lunch, a projection screen and microphone were available throughout the day for teams to informally announce insights and questions. Teams also used Dropbox, Google documents, GitHub, and Challengepost for online collaboration and analysis. Technical staff with knowledge about Ebola, disease modeling and data analysis helped guide participants with computational but not epidemiological knowledge toward interesting questions and around obstacles. In general, we found it useful for participants and technical volunteers with relatively unique areas of specialty to introduce themselves during the event so that teams knew to whom to refer questions which required special training or experience – fields covered by our participants included most commonly statistics, data analysis and computer science; the fields covered by our technical team included these areas as well as well as geography, artificial intelligence, epidemiology, theoretical ecology, and Ebola specifically. A logistics engineer specializing in disaster relief with Oxfam (who had assisted in relief efforts following the 2010 Haiti earthquake) also volunteered to provide additional context around emergency and crisis management processes.

Concluding the event. Teams presented their findings to each other and representatives from the Haitian Health Ministry at lunch on Sunday. The Haitian Health Ministry representatives reported that attending HackEbola (with Data!) was valuable not only because the presentations were relevant to their own emergency preparedness planning, but also because they saw how such an event could be used to quickly gather insights from complex data. The representatives also gave practical feedback to teams based on their experience in crisis response, medical treatment, and psychological-sociological factors. Through these presentations, teams also gained experience in conveying their results to humanitarian crisis experts.

2.4. Observations for future events

A data-analysis hackathon such as HackEbola (with Data!) lies in an interesting niche. Unlike broader hackathons, where participants come together to brainstorm (any, not necessarily involving data analysis) interesting solutions to a problem (such as HackEbola @NYU), HackEbola (with Data!) was focused around a very specific data analysis objective. For this kind of hackathon to be successful, one must understand the demographic that such an event tends to attract: most of the participants were data experts of some kind–computer scientists, machine learning experts, statisticians and biostatisticians. They had deep expertise in computational problems, statistical theory, and working with complex-structured data. However, very few had experience in geography, epidemiology, and crisis management. To structure an event around the strengths of this type of participant, we emphasize the need for:

- *Clear, appropriately scoped, and open-ended objectives.* Since participants are unlikely to know much about the specific domain, the challenge must guide participants toward the interesting questions. However, it should be sufficiently open-ended regarding the methodologies so that participants can apply their deep data expertise to the problem—a hackathon is not an appropriate format for a standard data analysis task with an obvious methodological approach. HackEbola (with Data!) centered around a concrete prediction and association-mining task, without specifying how the objective should be achieved.
- *Vetted, well-formatted data.* While data scientists are experts at working with complex data, they are not necessarily experts in other areas relevant to crisis management (e.g. geography). To leverage their strengths in data analysis, the data must already be easily machine-readable into tables (several teams contacted HackEbola (with Data!) in advance for the table schema). These data tables must have consistent indices on which to link them (ensured by qDatum.io). Finally, a one or two day event leaves little time for outside research and remedying data irregularities. Teams will often be unacquainted with alternate data sources, and even when alternate sources are easily available, there may not be time to process them. Thus, one must ensure that data of sufficient quality and quantity to address the hackathon objective is available prior to the event.
- *Technical staff.* In addition to standard event logistics (i.e., providing meals, microphones, and electronic outlets for laptops), it was important that technical staff with Ebola, crisis, and epidemiological expertise were present. These staff members helped familiarize teams with data pre-processing techniques, modelling assumptions, and alternative data sources that were relevant to the specific domain.
- *Attendance and investment by stakeholders.* For data experts unfamiliar with the specific domain, it is both extremely educational and motivating to have stakeholders present at the event. Having a representative from the Red Cross open the event helped make the challenge real; likewise, it was valuable for teams to be able to discuss their approaches with the Haitian Health Ministry representatives. Similar investment from other organizations would have greatly increased the impact of the event: at the end of the day, participants volunteered their time

because they believed that organizations would use the outputs of their analyses. While an event summary was circulated after the event among Statistics without Borders, the Red Cross, and Digital Humanitarian Network, the investment shown by attending the event increases motivation and can assist teams in prioritizing their analyses toward the most pressing and promising objectives.

3. Technical Outcomes from HackEbola with Data

3.1. Task

HackEbola (with Data!) focused on the following question: What factors may be affecting the regional and temporal evolution of the Ebola epidemic in West Africa? Specifically, teams were tasked to:

- Train a model or mine associations between the sets of cases and death time-series and covariates for data through October 1, 2014.
- Test that model (or the strength of the discovered associations) on data from October 2, 2014 to November 20, 2014.
- Report (a) how well the model performed (prediction accuracy) and (b) any factors that might be valuable in assisting with decisions to contain the outbreak. Teams were also asked to discuss limitations due to data quality and modeling assumptions.

3.2. Data

Most participants focused on publicly available datasets from the Humanitarian Data Exchange and the Open Humanitarian Data Repository. These included sub-regional time series of Ebola cases and deaths (34 regions across Liberia, Sierra Leone, and Guinea) from a variety of organizations, including the WHO and the respective national governments. The data also included sub-regional information such as the locations of Ebola treatment centers (sometimes with opening and closing dates), educational facilities, police stations, markets, and places of worship. Also provided where information about movement restrictions, population and food prices, and socioeconomic indicators including international wealth index scores; mean years of education for males and females; urbanicity; proportions of the population involved in various occupations; living quality indicators such as house size, available water, and electricity; and age proportions. These data were pre-processed by qDatum.io to have consistent administrative region identifiers across all of the tables. These data are available at:

http://www.qdatum.io/public-sources

Several teams shared processed versions of these data, as referenced in the Appendix.

3.3. Strategy

To assess what factors may be explaining inter-regional variation, most teams took the following approach:

- *Defining Dependent Variables.* Most teams processed the time-series in each region into a set of summary statistics, such as total cases, total fatalities, case fatality rate (CFR), transmission rate (TR), cases per capita, and fatalities per capita. A variety of assumptions were used to derive these summary statistics: simple counts, exponential and logistic models, and variations on SIR models.
- *Collecting Independent Variables.* While most teams focused on a standard of regional indicators, some teams built additional variables such as the effect of recent civil wars. One team incorporated a model of movement patterns into the standard SIR model using the number of roads between districts. The resulting model approximately consisted of having a set of SIR compartments for each region and having population flows

between (compartments of different) regions. Though the model had high prediction errors, forming the model in this way might assist with future efforts.

- *Analysis.* Teams used a variety of methods to compare the dependent and independent variables, including correlation computations, generalized linear regression, and other machine learning approaches. Teams hoped to discover factors that could potentially be modified by public health programs—such as movement restrictions and ETU locations. While a weekend was not a long enough time for an in-depth analysis, promising models or factors that seemed to be highly predictive could be used to initiate follow-up analyses.

More generally, some teams also investigated other factors, such as the change in food prices in the region and mapping accessibility to current Ebola treatment units (ETUs).

3.4. Main Findings

As published in many sources, all teams found that standard SIR models could fit national-level data well. However, data at sub-national levels contained many more irregularities. It is important to note that many teams did not find anything significant in the data, and even those trends that were statistically significant must be treated with extreme caution due to the biases in the data (more details in Limitations and Discussion section). Combined with the short nature of the event, which allowed for less time for thorough statistical analysis, in this section we only report the main trends and not specific correlation coefficients or p-values. For specific details reported by various teams, we instead point the readers to the hackathon gallery at:

http://hackebolawithdata.challengepost.com/

Individual Correlations. Following urbanicity, the most consistently reported covariate was age and country. Several teams also noted that the age distribution seemed to have an impact on the growth rate: rates were higher if the number of adults (20-60 years) was higher and lower among populations with more children (0-9 years) or seniors (over 60 years). This effect may be because the adult population is more mobile than young children or the elderly. Many teams also found that the country (or almost equivalently, the latitude) was one of the most dominant effects when predicting the growth rate, suggesting important country-specific distinctions.

When regressed individually, several teams found that urbanicity and level of education seemed to be positively correlated with case and fatality rates; most likely these variables may be a proxy for populations that are more likely to be better reported. Similarly, various teams found increased case and fatality rates for higher floor quality, cell phones, and flush toilets (and reduced rates for poor floors and poor toilets). While these teams also found increased case and fatality rates among sub-national regions with bad water and small houses, it seems that these correlation analyses are confounded by which areas have better reporting (i.e., those that are more urban and wealthy). These indicator data were also collected over a span of a decade prior to 2014; thus they may also be dated.

Combined Regressions. Teams that performed combined regressions also found that the urbanicity and education variables had the strongest effects. One team found that higher education had a protective effect, as did electricity. The Poisson regression employed by that team predicted observed cases after Oct. 1 with over 90% accuracy; they later presented a polished version of their results at a training session at the World Health Organization (WHO). However, teasing apart colinearities and confounds is challenging; for example, this team also found that having tap water had a negative effect.

In general, the use of a combination or overlap (e.g., in Venn Diagram terminology, the intersection) or comparison/contrast of team results in order to predict the most robust results which should then through follow-up be further checked for their potential utility to decision-makers (such as public health or crisis programs and responders) is meant to serve as an analogy to the public health strategy of combining multiple epidemiology models in order to make infectious disease control decisions; for example, Keeling and Rohani in their modern text of *Modeling Infectious Diseases In Humans and Animals*, 2008, (this book generally serves as a modern sequel to the classic summary text *Infectious Diseases of Humans* by Anderson and May) describe the possible advantages of combining multiple models and statistical methods and its role in the control of the foot-and-mouth disease 2001 United Kingdom outbreak.

Other directions. Some teams decided to perform other data analyses such as monitoring changes in food prices or mapping which areas are near ETUs. These are referenced in the Appendix.

3.5. Limitations and discussion

In modeling the Ebola outbreak at a sub-national level, teams were fundamentally limited by the quality of the data:

- *Irregularities in Case Data.* All participants reported many irregularities in the case data. Case data had many jumps reflecting the fact that the data indicated when cases were reported, not when cases occurred. In some regions, the number of deaths exceeded the number of cases—which could be explained by movement between regions or gaps in case reporting. Summing cases across regions did not produce numbers that matched national-level statistics. Cases were reported by five different sources with widely varying values; participants had to make difficult choices about which sources to use or how to combine sources. Different time periods had data at different levels of regional and temporal granularity.
- *Granularity of the Data.* Because of the many irregularities in the case data, most teams focused on deriving a few summary statistics of the sub-national time series (in particular, it was impossible to ascertain whether a spike or dip in cases was due to reporting or the effect of an intervention). Thus, the multi-month time series were collapsed into 34 regional data points.
- *Time of Covariate Collection, Missingness.* Teams focused on food prices noted that prices were often missing during the outbreak, and the dates when ETUs were opened or closed were also often missing. Covariates (socioeconomic status, education, etc.) were all collected prior to the outbreak, some as early as 2007.

Thus, the results above must be treated with extreme caution. In a detailed analysis after the event, one team showed how while models could be trained to fit the national-level data well, potentially artefactual variability in the case count trends in the sub-regions made it challenging to draw conclusions. In particular, small changes in training choices -- type of interpolation, choice of smoothing technique, and whether to train on incidence of new cases or cumulative counts -- resulted in large differences in the parameters. Teams found that changing some of the assumptions used to compute the dependent variables resulted in spurious correlations. One team provided several examples, including a positive correlation between education and case fatality rates. These effects are in addition to the challenges due to confounds (more urban areas are likely to have better reporting) and colinearities (more urban areas also have more educated citizens).

4. Conclusions

HackEbola (with Data!) created a considerable interest in the data science community to help address an important need: understanding how various regional factors might be affecting the temporal and geographic evolution of the epidemic. Even when advertised as a “small” event, it resulted in over 300 registrants, of which 189 participants downloaded data. The event demonstrated how a hackathon format could be used to test many approaches quickly: each team used different assumptions when pre-processing and modeling the data. However, in the end, teams were still limited by the quality of the data. In the future, “deep dive” events such as these might benefit from additional steps to check whether the data are amenable for such fine-grained analysis.

5. Acknowledgments

We would like to thank the crisis experts who consulted teams during the Hackathon: Colby Wilkason, Dr. Marie Ghislaine Adrien, and Christine Duchatellier Fowler.

We would also like to acknowledge our supplementary proofreaders: E. Marshall who has training and experience in emergency and first response care, and R.K. Warner who is educated in health sciences.

6. Authors

This report was written by F. Doshi-Velez (Harvard University) and Y. E. Marshall (Princeton University), and HackEbola with Data was partnership among staff and participants. Our staff were F. Doshi-Velez, Harvard Organizer. D. Lasry, Y.E. Marshall, R. Gong, G. Harling, Q. Wang, Harvard Technical Staff. M. Kapp, J. Ming, A. B. Bind, Harvard Site Staff. N. Reich, K. Gourgoulias, E. Ramos, Amherst Organizers. E. Leschem, I. Maltz, Data Management Team qDatum.io. J. Pollack, Statistics without Borders representative. Our contributing participants were: M. A. Testa, M. Su, S. Konate, J. Torres, and E. Savoia; G. Khimulya, D. Sampas, and R. G. Cutbill; Y. Gurmu, G. Harling, S. Vardhanabhuti, S. Chin, O. Patterson, L. Valeri, A. Ablorh, J. Bobb; D. Schoenfeld; M. Duanmu; C. Yang and S. Okuda; X. An, Y. Chang, J. Chen, L. Hou, and X. Li; Y. E. Marshall, S. Mahimkar, R. A. Zelaya, T. Fussell, and A. Low; G. Bridgman, Q. Truong, A. Amin, A. Jacobovits, and E. Jospe; C. Valentine and D. Roscoe; S. Wang and W. Liu; A. Low; G. Sabran, K. Altenburger, K. Fodouop, S. Wang, and M. Andere; and J. Winget.

7. Appendix

7.1. Processed data sets

Several teams created processed data sets for other teams to use. One team created a cleaned up set of data containing the date of the start of the epidemic in the region relative to the start of the epidemic, duration of the epidemic in that location, the total number of cases and the total number of deaths. They also estimated a parameter beta which is the log of number of cases over the duration (estimates the rate of spread):

https://www.dropbox.com/sh/kniy9s8rcuvxu1b/AACpzPnkpNu7IrQMdyTBRL_Va?dl=0

Others also shared data sets with gaps interpolated and files merged with indicators:

https://www.dropbox.com/sh/ief4x9yshmd6619/AADJgB49a_xQcwl9aYynv2Pua?dl=0

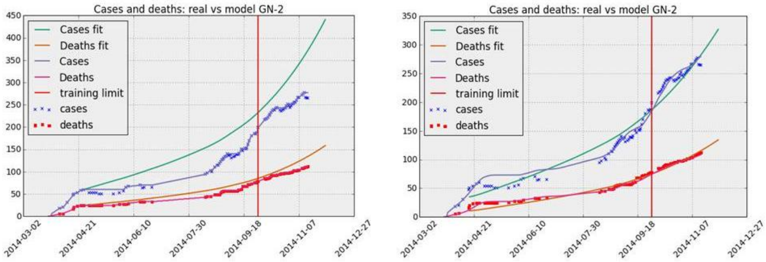
https://drive.google.com/folderview?id=0B915sMG77RJYblBTS2hhdURaMms&usp=sharing

Another team estimated the number of treatment beds per region and shared the results:

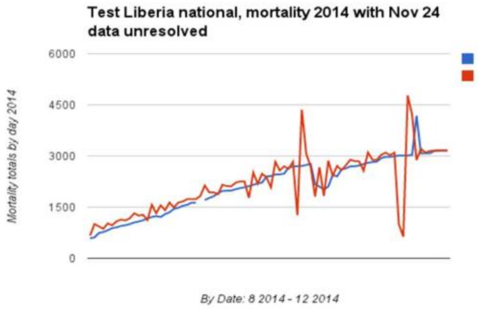
https://www.dropbox.com/sh/qtibq5tv2yful8l/AADP4tuhvD1HRYG3bLAorM_Ia?dl=0

7.2. Example Plots

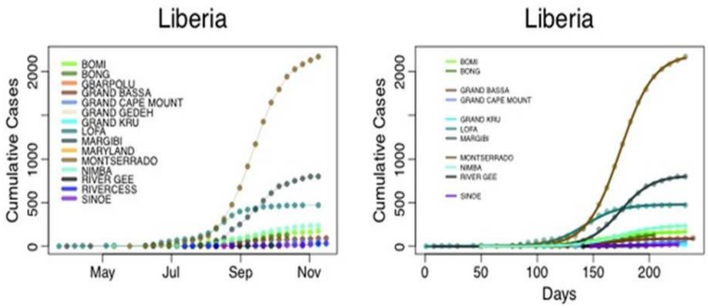
We avoided reporting specific values or including specific plots from the teams in the main document, as the analyses were computed over a single weekend. Here we include some example plots to show the kinds of analyses that were performed. The team of G. Sabran, K. Altenburger, K. Fodouop, S. Wang, and M. Andere noted a large difference between fitting the SIR model to smoothed and unsmoothed case data, as seen below for the case of Guinea. They also noted a wide variation in the shapes of case curves at the sub-national level.



The team of J. Winget computed the following fit to the SIR model of fatalities in Liberia; a key question was which data—from a variety of government and organization sources to use.



The team of Y. Gurmu, G. Harling, S. Vardhanabhuti, S. Chin, O. Patterson, L. Valeri, A. Ablorh, and J. Bobb created the following plot using a logistic curve to fit case time-series in Liberia. Similar plots for Guinea and Sierra Leone are in their ChallengePost entry.



Below is a representative Poisson regression analysis by the team of M. A. Testa, M. Su, S. Konate, J. Torres, and E. Savoia.

Table 7

GEE population-averaged model		Number of obs	= 522
Group variable:	sdr_id	Number of groups	= 34
Link:	log	Obs per group: min	= 2
Family:	Poisson	avg	= 15.4
Correlation:	exchangeable	max	= 44
		Wald chi2(4)	= 55.37
Scale parameter:	1	Prob > chi2	= 0.0000

(Std. Err. adjusted for clustering on sdr_id

Cases	IRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval
country_code2						
GN	.4020459	.2834323	-1.29	0.196	.1009717	1.60085
LR	.8172721	.2061169	-0.80	0.424	.4985316	1.33980
age09	.8213987	.054686	-2.96	0.003	.7209147	.935888
electr	.9660433	.0137967	-2.42	0.016	.9393772	.993466
_cons	.0535935	.1235697	-1.27	0.204	.0005841	4.91715
ln(populat~n)	1	(exposure)				

7.3. Additional analyses

Food Prices. A few teams also looked at the changes in food prices during the outbreak. Teams also found that food prices, where available, had stayed mostly stable through the outbreak. Specifically, the changes in prices during 2014 appeared to be within the already broad price variability due to other factors in the region.

Visualizations and Geography. Finally, a few teams focused on exploring the data through visualizations and geographic analysis. In particular, A. Low created a map showing locations that were within an hour of an ETU: