Using Satellite Imagery to Predict Health

Jeremy Irvin
Department of Computer Science
Stanford University
Stanford, CA 94305
jirvin16@cs.stanford.edu

Dillon Laird
Department of Computer Science
Stanford University
Stanford, CA 94305
dalaird@cs.stanford.edu

Pranav Rajpurkar [*]
Department of Computer Science
Stanford University
Stanford, CA 94305
pranavsr@cs.stanford.edu

Abstract

Obtaining quality data on population health statistics in developing countries is currently an expensive and infrequent process. Policy makers must rely on this sparse data in order to decide where to provide aid. In the same direction as recent works, we make efforts to automate this data collection through the use of remote sensing with deep learning. We use Convolutional Neural Networks (CNNs) to directly predict poverty and malnutrition from satellite imagery. We find that directly predicting malnutrition is difficult, but our model can successfully classify impoverished regions with relatively high accuracy.

1. Introduction

Organizations such as UNICEF, the Bill and Melinda Gates foundation, and many more typically use surveys to assess where to allocate resources for humanitarian efforts in third world countries. The surveys are costly and potentially dangerous to collect, resulting in large temporal gaps between surveys and large spatial gaps between survey locations. This makes providing aid a very slow, sparse, and expensive process. Achieving more fine-grained health outcome predictions is invaluable to these foundations.

Remote sensing data has promising potential to predict health-related statistics [1, 2, 3]. This data contains a plethora of valuable information, from vegetation and land cover to city quality and development. Brown et al. demonstrate through case studies that remote sensing data can help predict environmental change that may directly affect local economies and food production [1]. Both climate change and short term weather patterns have been shown to be predictive of health in Africa and Nepal [4, 5]. Recent work describes how land surface phenology directly affects food security in Niger and Kenya [6].

We hope to expedite and cheapen the efforts by using satellite images and auxiliary data to predict different indicators of poor health, such as poverty, malnutrition, and malaria. This would allow these organizations to obtain much more frequent and cost effective global predictions of human health, including areas in which they are unable to obtain survey data. We congregate data from numerous Demographic Health Survey (DHS) surveys and combine them with Google Maps satellite images. We then use this labeled dataset to train a deep CNN to predict population health statistics from satellite images alone.

2. Related Work

Deep learning has been successfully applied to satellite data. CNNs have been used to classify land use via satellite images [7]. They use classes such as agriculture, buildings, forest, golf course, and parking lot, and find that using pretrained models performs the best. In a similar direction, CNNs have been used to perform the binary classification building vs. non-building on satellite imagery [8]. This works finds that using a fully convolutional neural network works better for satellite data: the model makes dense predictions by enforcing that the output is a result of a series of convolutions only. Another recent work uses ortho multispectral satellite imagery to effectively perform per-pixel classification of vegetation, ground, road, building, and water [9].

[*] Non-231N

Satellite data has also been used to predict crop yield in the United States and Africa [10, 11]. You et al. combine CNNs and Gaussian processes to predict county-level soybean production in the United States. A similar technique is applied on an area in western Kenya to predict crop yield on corn for small half-acre to one-acre lots [11]. While estimating crop yield might be a good indicator of health, we do not explore this task in this work.

There have been previous efforts to forecast malaria outbreaks using satellite imagery [12, 13, 14], but to our knowledge no work has used any form of deep learning. These works use statistical techniques to predict the distribution of malaria vectors (species of mosquitos which commonly carry malaria) from satellite imagery. We believe remote sensing data has potential to predict malaria outbreaks as well as other diseases. We do not experiment on this task in this work, however, as we could not obtain the data in time.

Remote sensing data provides a cheap and efficient way to measure socioeconomic indicators which are especially relevant for monitoring third word countries. One of the large issues with this task is lack of data, which has recently been circumvented by using transfer learning and proxies [15, 16]. Their idea is to use a pretrained model on ImageNet for extracting low-level features (like edges, for example) and then use nighttime light intensity data, a much more rich data source, as a proxy for economic activity to further train the model. Another approach is to use mobile phone data to track poverty [17] [18]. Unfortunately it relies on proprietary datasets from mobile phone networks which may not always be available. Our model trains directly on image, health label (poverty or malnutrition) pairs without the use of any proxies.

3. Dataset

We use Demographic Health Survey data [19] which contains more than 300 surveys in over 90 countries, where each survey contains health information from thousands of households. Figure 1 shows the distribution of households within the city of Harare, Zimbabwe in the DHS dataset. The households are grouped into clusters, and each cluster is assigned a latitude-longitude pair of its approximate centroid. However, in order to ensure confidentiality, the centroid is displaced up to 2km in urban areas and 5km in rural areas. This requires us to examine 10km by 10km regions for each cluster.

For poverty prediction, the particular statistic we use is the wealth index which is a "composite measure of a household's cumulative living standard" [20]. This score is computed from various features, for example the presence of televisions, materials used to construct the house, access to water, and sanitation. The wealth index is then divided into five buckets: poorest, poorer, middle, richer and richest. Examples of regions around a poorest household and a richest household are shown in Figure 2.

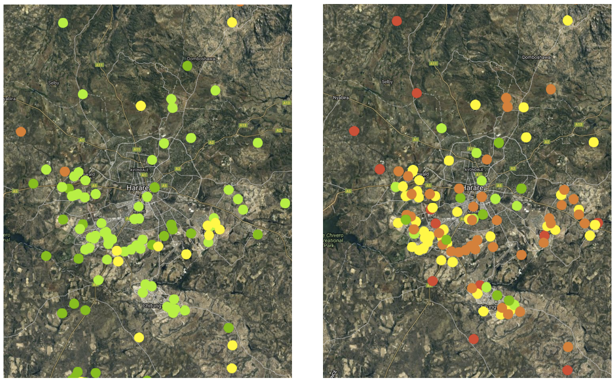

Figure 1. Households from DHS in Harare, Zimbabwe. The left image shows the households colored with their wealth label: red, orange, yellow, yellow-green and green correspond to poorest, poorer, middle, richer and richest households. The right image shows the households colored with their WAZ label: red, orange, yellow, yellow-green and green correspond to less than -1.5, -1.5 to -1, -1 to -0.5, -0.5 to 0 and 0 to 0.5 standard deviations around the mean.

For malnutrition prediction, the DHS dataset contains three relevant measures: height-for-age (HAZ), weight-for-height (WHZ), weight-for-age (WAZ) z-scores. HAZ reflects health and nutritional conditions of development during and after birth. Low HAZ (stunting) is a strong indicator for long-term nutritional deficiency and/or repeated illness. WHZ is a short-term measure and is more sensitive to recent severe events. Low WHZ (wasting) is substantial weight loss, often due to recent disease or lack of food. WAZ is combination of HAZ and WHZ. Low WAZ (underweight) measures both chronic and acute malnutrition. These z-scores are computed from global population statistics. We bucket these scores into six classes: less than 1.5 standard deviations below the mean, between 1.5 and 1, between 1 and 0.5, between 0.5 and 0, between 0 and 0.5 standard deviations above the mean, and more than 0.5 standard deviations above the mean.

We can see from Figure 1 the relationship between the poverty and malnutrition labels. While poverty is fairly consistent across the city, malnutrition is much more sporadic making it more difficult to predict. We encounter this issue when experimenting with different models as well.

For clusters that contain more than one household, we take the median label (wealth index bucket for poverty prediction, WAZ bucket for malnutrition prediction) over all households in that cluster and use that as the label for the cluster. Additionally, we only use DHS cluster data which was obtained in 2010 or later as the satellite images we have access to are much more recent. [1]. This results in

[1] We plan to obtain temporally correct images in the future.

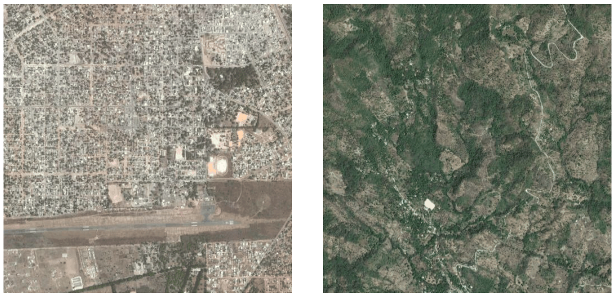

Figure 2. Example of a rich area (left) and a poor area (right).

around 31 thousand unique clusters clusters from over 45 different countries in Africa, Central America, South America, and South Asia. We split these clusters into training (80%), validation (10%), and test (10%). For poverty prediction, the class distribution of the training set is 18.8% poorest, 24.5%poorer, 21.6% middle, 18.3% richer, 16.5% richest. For malnutrition prediction, the class distribution of the training set is (in the same order as presented in the previous paragraph) 11.4%, 21.4%, 30.0%, 21.5%, 9.8%, 5.8%.

We use Google Maps Static API to pull satellite data of a 10km box surrounding each cluster. Since we are limited in the size of images we can pull and also the number of images we can pull due to throttling we test several different zoom levels. Each zoom level defines the resolution of the image. Zoom levels are between 0 (where the entire world is visible on the map) and 21 (where streets and buildings can be seen). We experiment with three different zoom levels: 13, 14 and 15. Each of these zoom levels correspond to about 20, 10 and 5 meters per pixel in resolution respectively when using 512×512 pixel sizes. At zoom level 13 we capture the entire 10×10 km square in a single image, while at zoom level 15 we use sixteen images each covering 2.5×2.5 km squares. To reduce the number of required images when using 2.5×2.5 km squares, we do not pull the tiles corresponding to the corners of the 10×10 km region, which results in twelve images per cluster. An example of this tiling is shown in Figure 3. We found that higher zoom levels improved the performance of the model. Unfortunately transitioning to even higher zoom levels would require more resources, or would force us to substantially subsample the regions, potentially hurting the models performance. We use zoom level 15 in all of the experiments and use twelve 2.5×2.5 km squares per cluster.

While the Google Maps API provides us with quick easy access to satellite data it is not perfect. Some regions did not contain data with the correct zoom level. For example the highest zoom level for some regions was 13 instead of 15. Cloud coverage and merged images was also a small issue. Certain regions were composed of several satellite images stitched together, sometimes at different zoom levels, which can be difficult for the algorithm to work with. We estimate these flaws represent less than 5% of the data. Some examples are shown in Appendix B.

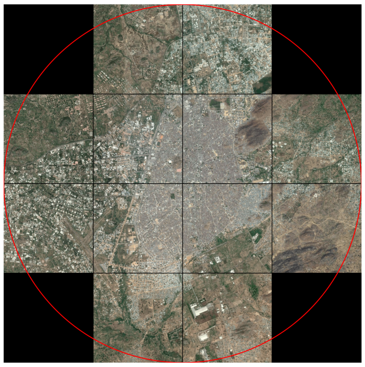

Figure 3. Example of tiling a single cluster region into twelve higher resolution images. The red circle indicates the region where the households may lie.

4. Methods

The basic structure of the model is a single CNN that is run over each tile (twelve tiles where each tile is a 2.5×2.5 km square as discussed in Section 3). A representation from each image is then combined to produce a final output. We explore several different convolutional architectures. The simplest consists of three blocks of convolution, max pool, ReLU followed by two fully connected layers.

We also experimented with a residual network (ResNet) [21]. We chose ResNet because of its good performance on common image classification datasets like ImageNet. ResNet models have high accuracy relative to their number of operations and number of parameters compared to other models such as VGG and Inception [22]. This is especially beneficial for our use case since we do not have a large amount of training data which makes it more difficult to train networks with a large number of parameters. ResNets use shortcut connections to improve gradient flow. More precisely it introduces the following function:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$$

where $\mathbf{x}$ is the input into the residual block and $\mathcal{F}$ consists of a convolution, batch normalization, ReLU, convolution, batch normalization. These types of connections allow the network to grow very deep without incurring common optimization issues associated with deep networks. Figure 4 illustrates a ResNet type block which has shown empirical

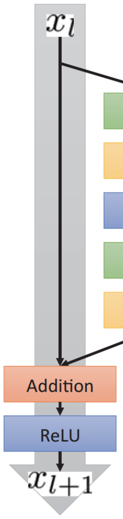

Figure 4. Example of a standard ResNet block. Figure taken from [23].

success in other works. We use this block in our ResNet models.

We test several 18-layer ResNets. We train one end-to-end and three others pretrained on ImageNet, where we tune the final three, two, and last blocks (as well as the final fully connected layer) on our dataset. Because ImageNet consists of 224 × 224 images and the images in our dataset are 512 × 512 we randomly crop a subsection of the image to run the pretrained ResNet on.

One of the issues with running the model over twelve images is that each image should not contribute the same amount to the final output. For example, if one image contains a town and the other images contain brush, only the image containing the town should contribute to the final output since the other images do not contain any signs of people. Although the network should ideally learn to handle this, we explore several methods to better equip the model to combat this, described below.

In order to combine the tiles of the cluster to make a single prediction, we explore different combining schemes: spatial vs. scores and max vs. mean. In the spatial scheme, we run each tile through the same CNN to compute a spatial representation per each tile. These representations are then combined through either a max or mean operation. In the scores scheme, we run the network on each tile, but additionally flatten the spatial representations and run those through a two-layer, fully-connected neural network to produce scores for each tile. These scores are then combined through either a max or mean operation. Figure 5 illustrates the general combining method for four tiles. For the scores scheme we have

$$c_t = CNN(x_t), \quad t = 1, \ldots, 12$$

$$o_t = Average([c_1, \ldots, c_{12}]), \quad o_t = Max([c_1, \ldots, c_{12}])$$

where x_t is one of the images, c_t is the output of CNN for image t and o_t is the combined representation which is then fed through another fully connected layer before it is fed into a softmax.

In addition to the max and mean operations, we explore combining the output representations from the CNN with a recurrent neural network. The hope is that the model could learn to retain information from images with significant features and forget ones without relevant information. As above each CNN produces some output c_t which is then fed into a long short-term memory network (LSTM)[24] to produce a final output,

$$o_t = LSTM(c_t, h_{t-1}), \quad t = 1, \ldots, 12$$

$$logits = Linear(o_{12}).$$

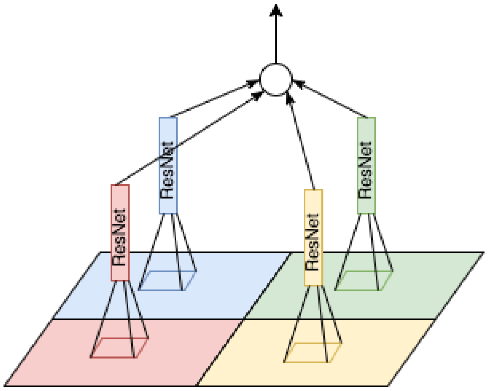

Figure 5. Each of the tiles are fed through the network (ResNet shown here) to compute either a spatial representation or a score vector. The white node represents the combination operation using either a mean or max operation or an LSTM.

5. Experiments

We experiment with three main models as described in Section 4 with many different hidden layer sizes, kernel sizes, model depth, and more. The validation and test accuracies of the best models on poverty prediction are shown in Table 6.

We only compare spatial and score combining schemes on the non-ResNet models as those models output a flattened, nonspatial representation. We find that combining the spatial representations in these models performs marginally better than combining the scores (by around 3% on average). Additionally, we find that there is minimal difference between the max and mean combining operations, with mean slightly outperforming max (by around 1.5% on average). We also test the LSTM combine scheme on the pretrained ResNet model.

	Poverty	
Model	Valid	Test
3 Conv/Max Pool/ReLU, 2 FC	34.541	32.681
ResNet Fully Trained	34.253	32.906
ResNet Pretrained, Mean	**44.131**	**43.489**
ResNet Pretrained, LSTM	34.092	32.970

Figure 6. Experimental results on the validation and test sets. Human gets 32% accuracy on a sample of 100 images.

	Malnutrition	
Model	Valid	Test
ResNet Pretrained, Mean, 0.0	28.726	21.738
ResNet Pretrained, Mean, 0.1	33.927	22.854
ResNet Pretrained, Mean, 0.2	29.773	21.982
ResNet Pretrained, Mean, 0.3	33.089	21.633

Figure 7. Experimental results on the validation and test sets for malnutrition prediction. Model column displays the model, combine operation, and dropout probability.

We then apply the model which performs the best on poverty prediction (ResNet Pretrained with mean combining operation) to the malnutrition prediction task. The validation and test accuracies of this model tuned on malnutrition labels with different values of dropout probability are shown in Figure 7.

In all experiments, we use Adam with a learning rate of 0.0003, dropout probability ranging from 0.1 to 0.4, and L1 weight decay ranging from 0.0001 to 0.1. We use batch sizes of 6, 12, and 32 depending on the model size, and train for 50 epochs (the model begins to overfit within this period in each of the experiments). We select the model which achieves the best validation accuracy to run on test.

6. Discussion

Our best model does well predicting poverty, achieving around 44% accuracy on both validation and test. It is interesting to note that the pretrained model performs the best, which somewhat challenges our intuition: although using pretrained models on smaller datasets tends to perform better than fully trained models, we did not expect many of the feature extractors learned from the ImageNet dataset to be useful in this setting. Satellite images are much different than images in the ImageNet dataset. However, we see in Figure 9 that the model can detect infrastructure such as buildings and roads and, to a lesser extent, it can discern the outline of farm plots. This is evidence that the first layer filters of a ResNet model pretrained on ImageNet can be useful for satellite images as well.

Figure 8 illustrates that most of the models mistakes are off by a single class. For example, the model often confuses poorer and poorest and has a very difficult time discerning poorer and middle. These are reasonable mistakes; our human tester also struggled on these exact cases. We suspect that higher resolution data could help mitigate these errors as more fine-grained features such as the quality of rooftops, presence of cars, etc. could make the task easier. It is worth noting that the best model obtains a top-2 accuracy 72.417% on validation and 72.996% on test, revealing that the model does very well in predicting the correct class in the top 2.

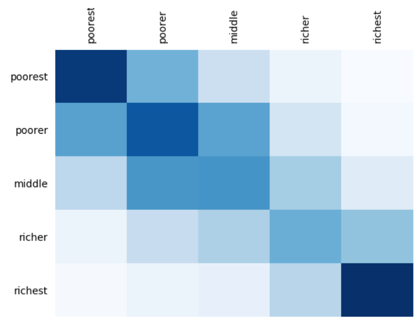

Figure 8. The confusion matrix for the ResNet pretrained on the validation set. The y-axis is the true label and the x-axis is the predicted label.

Our models for malnutrition prediction achieve decent validation accuracy but fail to generalize well to the test set in every experiment, as shown in Figure 7. We hypothesize the poor performance on malnutrition is mainly attributed to two reasons. First, the data for malnutrition is noisy. The WAZ scores (and other malnutrition metrics) are dependent on other factors, such as genetics, which may not be perfect indicators of poor health. Moreover, malnutrition is known to be well-correlated with poverty [25], but we find that the wealth and WAZ scores have a Pearson correlation coefficient of 0.17. Second, we hypothesize that there are very few direct features present in satellite data that are predictive of malnutrition, and if they do exist, they likely appear at a much higher zoom level. We hypothesize that this means there is little signal in the data. We believe the model is essentially overfitting the training set (and validation set since models are selected by validation score) and thus the model cannot learn features which generalize. Note that for this reason, increasing dropout does not relieve this issue, which is verified by our experiments illustrated in Figure 7.

Predicting poverty can be more easily performed with satellite images than with surveys as shown in this work as well as others [16]. This is because physical features that are correlated with poverty such as infrastructure can be clearly observed from satellite images. However, survey data still provides more accurate statistics. We believe this gap can be closed with future work. Moreover, poverty is one of the most influential risk factors for ill health [26]. Therefore effective poverty prediction may serve as an invaluable feature for predicting other health issues as well.

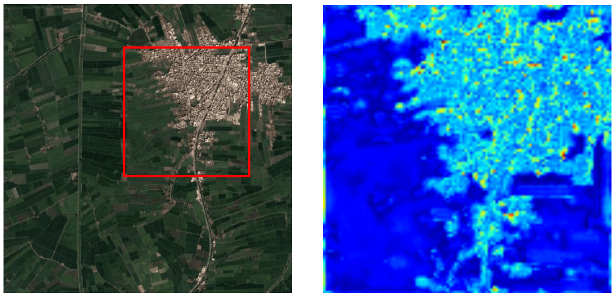

Figure 9. The input image (left) is randomly cropped to a 224x224 image to feed into the pretrained ResNet model. Visualization of activations (right) of the first layer of a pretrained ResNet model tuned on our datset.

7. Conclusion and Future Work

We show that satellite images can be used to predict poverty without the use of proxies. More notably, we demonstrate that models can perform well with a relatively small amount of data by using pretrained architectures. While not as accurate as survey data, we can quickly and cheaply obtain a fine-grained map of poverty levels for certain areas.

We hope to test models which are more selective about where to look in the satellite image, rather than randomly sampling a location. This could be accomplished by sending a downscaled version of the larger satellite image through an LSTM which outputs a location which is then used to feed an image at that location through a CNN, and then finally feed that output back into the LSTM and repeat. This could be trained with reinforcement learning. We also want to obtain higher resolution and higher quality satellite images which are temporally correct. The Appendix B illustrates a few of the issues with the current dataset. We believe the performance of these models can be greatly improved with the above changes and additions.

Moreover, we hope to extend this work to predict other health outcomes such as malaria, which are more correlated with physical features like wetlands, bodies of water, quality of rooftops, and highly dense populations. We are in the process of obtaining malaria data for many countries. We believe satellite images have promising potential to predict these health-related outcomes and hope to greatly further the progress in this direction.

Acknowledgments

Thank you to Pranav Rajpurkar for his frequent help on the project throughout the quarter, Dr. Eran Bendavid for proposing the project and providing the DHS data, and Andrew Ng for his helpful suggestions.

Appendix

A. Cross Entropy Loss and Accuracy Curves

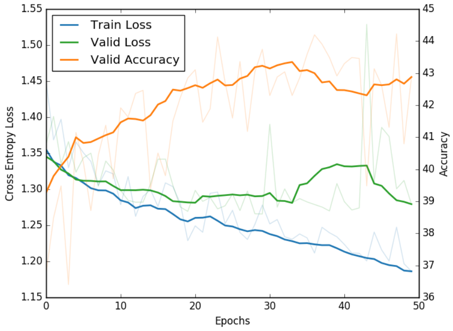

Figure 10. Training and validation cross entropy loss curves as well as the validation accuracy of the best pretrained ResNet model on poverty prediction.

We plot the cross entropy training and validation loss curves as well as validation accuracy of the best pretrained ResNet model on poverty prediction in Figure 10. There is a fair amount of variance in the loss. We believe this is due to label noise in the dataset, some of which may be due to problems described in Appendix B.

B. Problems with Satellite Data

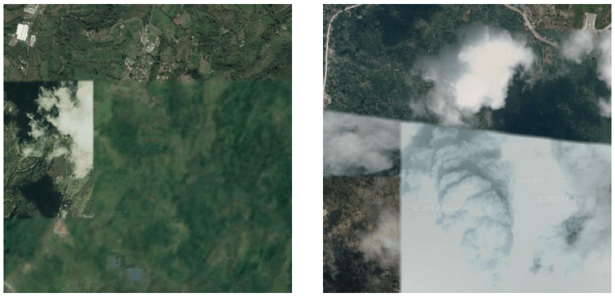

Figure 11. Two bad images collected from the Google API. The one on the left contains multiple zoom levels while the one on the right contains different crops with clouds

While scraping satellite data from the Google API relieves a lot of difficulties associated with satellite imagery it does not eliminate all of them. Figure 11 shows some of the bad images found in the dataset. Combining satellite data from different providers and keeping the images looking nice and consistent can be difficult. Fortunately, we estimate that these bad images represent a only a small portion of the dataset.

References

[1] M. E. Brown, K. Grace, G. Shively, K. B. Johnson, and M. Carroll, "Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change," *Population and environment*, vol. 36, no. 1, pp. 48–72, 2014.

[2] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE 2016*, vol. 104, 2016.

[3] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing: Challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[4] M. Kudamatsu, T. Persson, and D. Strömberg, "Weather and infant mortality in africa," 2012.

[5] P. Mulmi, S. A. Block, G. E. Shively, and W. A. Masters, "Climatic conditions and child height: Sex-specific vulnerability and the protective effects of sanitation and food markets in nepal," *Economics & Human Biology*, vol. 23, pp. 63–75, 2016.

[6] M. E. Brown, K. M. de Beurs, and K. Grace, "Global land surface phenology and implications for food security," in *Land Resources Monitoring, Modeling, and Mapping with Remote Sensing*, pp. 353–363, CRC Press, 2015.

[7] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv preprint arXiv:1508.00092*, 2015.

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[9] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sensing*, vol. 8, no. 4, p. 329, 2016.

[10] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep guassian process for crop yield prediction based on remote sensing data," *AAAI Conference on Artificial Intelligence*, 2017.

[11] M. Burke and D. Lobell, "Satellite-based assessment of yield variation and its determinants in smallholder african systems," *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[12] D. J. Rogers, S. E. Randolph, R. W. Snow, and S. I. Hay, "Satellite imagery in the study and forecast of malaria," *Nature*, vol. 415, no. 6872, p. 710, 2002.

[13] K. Pope, P. Masuoka, E. Rejmankova, J. Grieco, S. Johnson, and D. Roberts, "Mosquito habitats, land use, and malaria risk in belize from satellite imagery," *Ecological Applications*, vol. 15, no. 4, pp. 1223–1232, 2005.

[14] M. C. Thomson, S. J. Connor, U. D'Alessandro, B. Rowlingson, P. Diggle, M. Cresswell, and B. Greenwood, "Predicting malaria infection in gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results.," *The American journal of tropical medicine and hygiene*, vol. 61, no. 1, pp. 2–8, 1999.

[15] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *arXiv preprint arXiv:1510.00098*, 2015.

[16] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[17] L. Hong, E. Frias-Martinez, and V. Frias-Martinez, "Topic models to infer socio-economic maps," *AAAI Conference on Artificial Intelligence*, 2016.

[18] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[19] "The dhs program." `http://www.dhsprogram.com/data/`.

[20] "Dhs program wealth index." `http://www.dhsprogram.com/topics/wealth-index/Index.cfm`. Accessed: 2017-06-06.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512:03385*, 2015.

[22] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[23] W. Shang, J. Chiu, and K. Sohn, "Exploring normalization in deep residual networks with concatenated rectified linear units.," in *AAAI*, pp. 1509–1516, 2017.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] Unicef *et al.*, *Strategy for improved nutrition of children and women in developing countries*. Unicef, 1990.

[26] M. Pena and J. Bacallao, "Malnutrition and poverty," *Annual Reviews*, vol. 22, 2002.