Machine Learning and Geospatial Approach to Targeting Humanitarian Assistance Among Syrian Refugees in Lebanon

Ziqi Xu, Arrhan Bhatia, Cassi Chen, Anushka Mazumdar, Angela Lyons, Aiman Soliman
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

INTRODUCTION:

- **Syrian Refugee Crisis**: 30% of Lebanese Population is made up of Syrian Refugees as a result of the Syrian Civil War [1]
- **Ineffective Aid Distribution**: Ineffective humanitarian aid distribution due to the limited supply and large undocumented refugee population

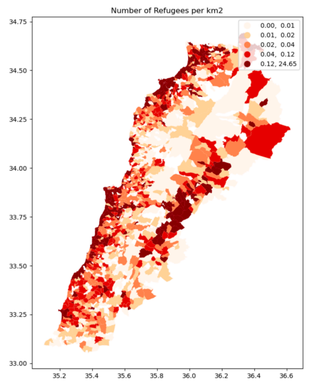

Figure 1: Map of refugee density distribution in Lebanon

GOAL:

- Predict where the largest refugee populations sit and key factors influencing the spatial distribution of Syrian refugees
- Develop an accurate and stable model.
  - **Accuracy:** How close model predictions are to outcomes. Used various model metrics to evaluate.
  - **Stability:** How robust model is to different sets of data. Compared model performance across data from different years.

LITERATURE REVIEW:

- Developed a conceptual model of refugee movement in Lebanon:
  - Read 30+ research papers concerning the migration of Syrian refugees to Lebanon
- Identified a large list of factors segregated into five broad groups:
  - Agricultural
  - Political
  - Economic
  - Social
  - Geographical
- Key predictors identified in literature review align with most important features identified by our models indicating the reliability of our model

METHOD:

Data Collection:

- **Credible Sources:** Collected data from organizations such as:
  - United Nations High Commissioner for Refugees (UNHCR)
  - United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA)
  - The Vulnerability Assessment of Syrian Refugees in Lebanon (VASyR)

Feature Generation:

- **Clean:** Removed errors, outliers, and null values from the individual raw datasets
- **Generate:** Mapped raster data and generated useful features such as distance to or sum of points
- **Combine:** Merged individual feature datasets into one final dataset to be used for model development

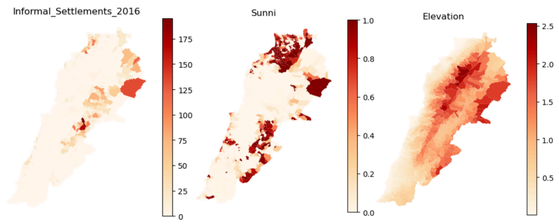

Figure 2: Maps of features generated using raw datasets

Preprocessing:

- **Normalization:**
  - Log Transformation
  - Standard Scaler
  - Proximity Interpolation
- **Multicollinearity:**
  - Correlation Coefficient
  - VIF

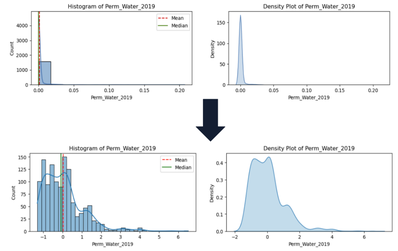

Figure 3: Impact of proximity interpolation on data skew

Challenges:

- Zero-inflated features
- Extraction of refugee, religion data from maps into CSV
- Inconsistency in cadaster names across dataset
- Lack of cadaster level data

Figure 4: OCR method

Figure 5: Interpolation method

Figure 6: Disaggregation method [2]

MODEL EVALUATION:

Exploration:

- **Models:**
  - Random Forest
  - Multi-Layer Perceptron (MLP)
  - Gradient Boosting
  - Support Vector Machine (SVM)
  - K-nearest neighbors (KNN)
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
- **Hyperparameter Tuning:** Utilized grid search to optimize model performance

Evaluation:

- **Metrics:** Evaluated model performance on a variety of metrics:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Coefficient of Determination (R-squared)
  - Mean Absolute Percentage Error (MAPE)

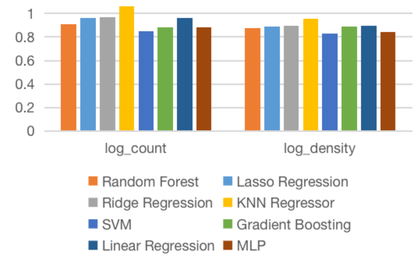

Figure 7: Model Performance MSE Result of log model

- **Residuals:** Utilized residual plots to assess accuracy, robustness, and reliability of predictions.

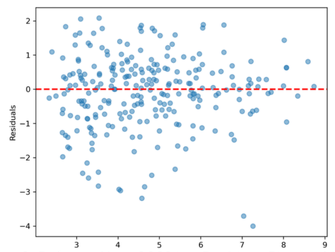

Figure 8: Refugee Num 2022 log Residual Plot by MSE

Results:

- Random Forest performs best from 2018 to 2022 on average
- Feature Importance:

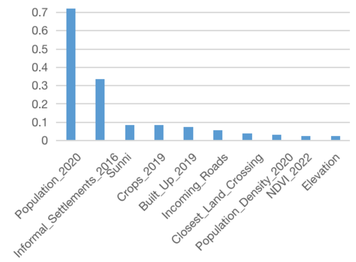

Figure 9: Feature Importance 2022 of log model

FUTURE WORK:

Current Model:

- Desegregate economic data from district to cadaster level

Other Models:

- Spatial Regression Model
  - Account for spatial autocorrelation
- More Complex Neural Networks
  - Enhance predictive power

CONCLUSIONS:

Our research highlights the effectiveness of using machine learning models to predict the distribution of Syrian refugees in Lebanon. By utilizing data from credible sources and thorough preprocessing to clean our raw data and improve model performance, our Random Forest model consistently performed best from 2018 to 2022. Our average percentage error on our best model was 23.5%. Furthermore, key predictors from the literature align with our model's most important features. These key predictors will greatly help humanitarian organizations better understand where Syrian refugees are located and thereby improve their aid distribution.

ACKNOWLEDGEMENTS

Thank you to the SPIN internship program and NSF for funding the REU FoDOMMaT. We also thank our mentors for providing us guidance throughout the program: Dr. Lyons, Dr. Soliman, Dr. Kass-Hanna, Yifang Zhang, Deepika Pingali, and David Zhu. We thank Anna Baskins, Nishk Patel, Sona Krishnan, Ishaan Salaskar, and Sarvagnya Vijay for their previous work. We also thank Illinois Computes for enabling access to NCSA's computation resources.

We appreciate the various organizations providing open-source geospatial data, including UN OCHA, UNHCR, WorldPop, Copernicus, Earth Observation Group, Uppsala Conflict Data Program, International Steering Committee for Global Mapping, USGS, Searates, Uppsala Conflict Data Program, VASYR and L'Orient Today.

REFERENCES:

[1] UNHCR Lebanon. UNHCR Global Focus. (n.d.). https://reporting.unhcr.org/operational/operations/lebanon#:~:text=The%20Government%20of%20Lebanon%20estimates,by%20the%20end%20of%202022.

[2] Rohan Arambepola, Tim, Nandi, A. K., Gething, P. W., & Cameron, E. (2021). A simulation study of disaggregation regression for spatial disease mapping. Statistics in Medicine, 41(1), 1–16. https://doi.org/10.1002/sim.9220

ILLINOIS