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

Arrhan Bhatia & Ziqi Xu, SPIN Interns
Cassi Chen & Anushka Mazumdar, REU FoDoMMaT Fellows
Angela Lyons & Aiman Soliman, NCSA Mentors

ILLINOIS

NCSA | National Center for
Supercomputing Applications

Motivation

Syrian Refugee Crisis:
- Ongoing civil war in Syria (Since 2011) has lead to large influx of Syrians refugees in Lebanon
- 30 % of the population in Lebanon consists of Syrian refugees

Ineffective Aid Distribution:
- High demand but insufficient supply of humanitarian aid
- Large undocumented refugee population

Our Goal:
- Help humanitarian organizations predict where the largest refugee populations sit and key factors influencing the spatial distribution of Syrian refugees
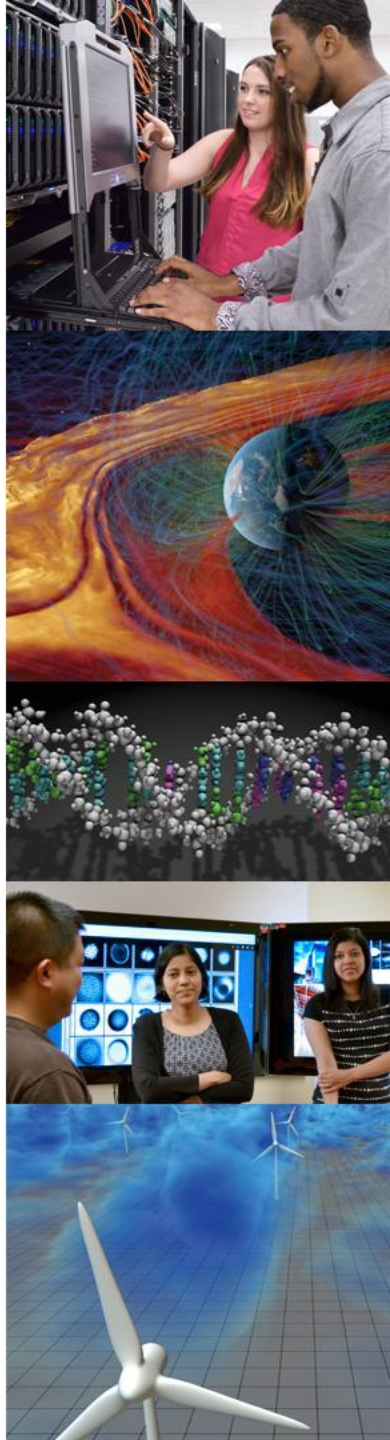
Previous Work

Summer 2023 REU/SPIN:

- Project started in Summer 2023
- Built foundation for data collection and analysis
- Provided preliminary results on a smaller set of features
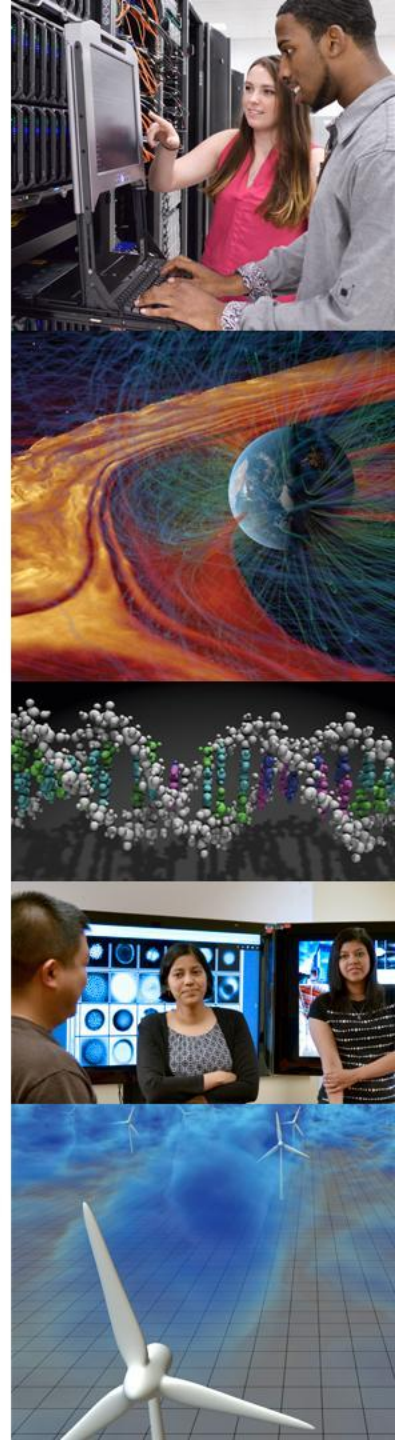
Weakness:

- Features were hypothesized
- Lack of literature review

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, and Geographical

ILLINOIS NCSA

Data

Data Collection

Agricultural & Political:

- Land Elevation
- Quality of vegetation
- No Camp Policy

Economic:

- Labour Market
- Availability of loans
- Financial stress
- Income Opportunity

Social:

- Healthcare
- Syrian refugee collective centers
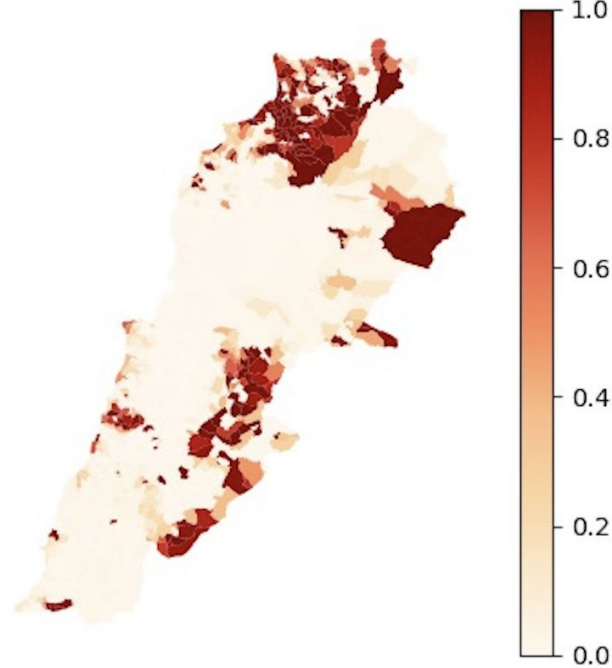- Conflict Counts
- Religious distribution

Geographical:

- Refugee population and population in each cadaster
- Latitude and Longitude
- Palestinian Refugee distribution
- Access to roadway and waterways
- Distance to nearest Syrian border
- Nightlight
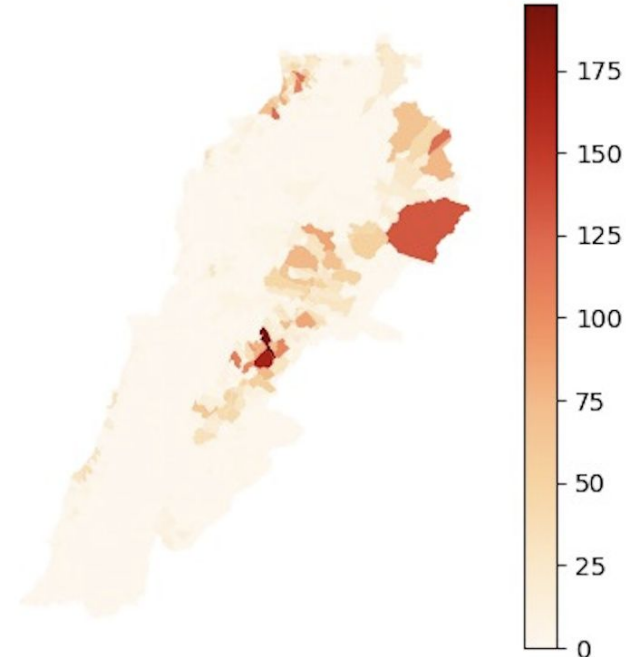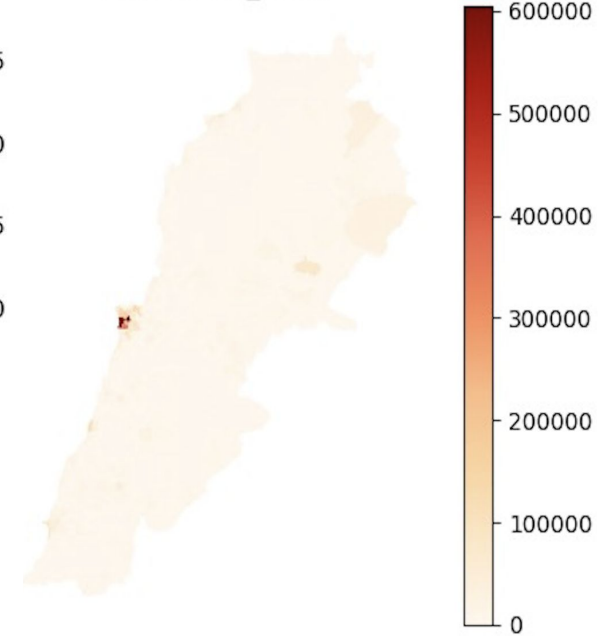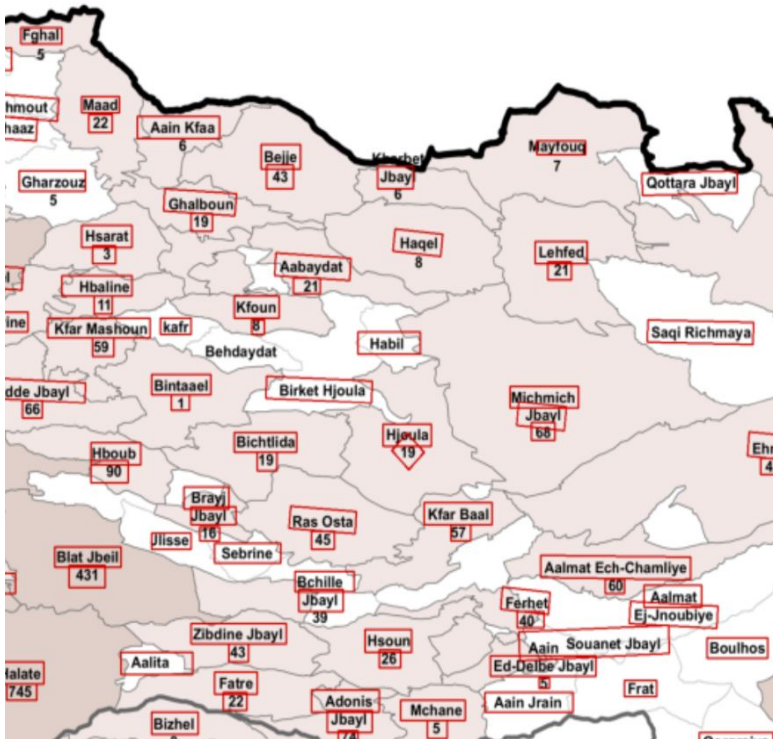- Average fraction of land area coverage

Some Data Examples
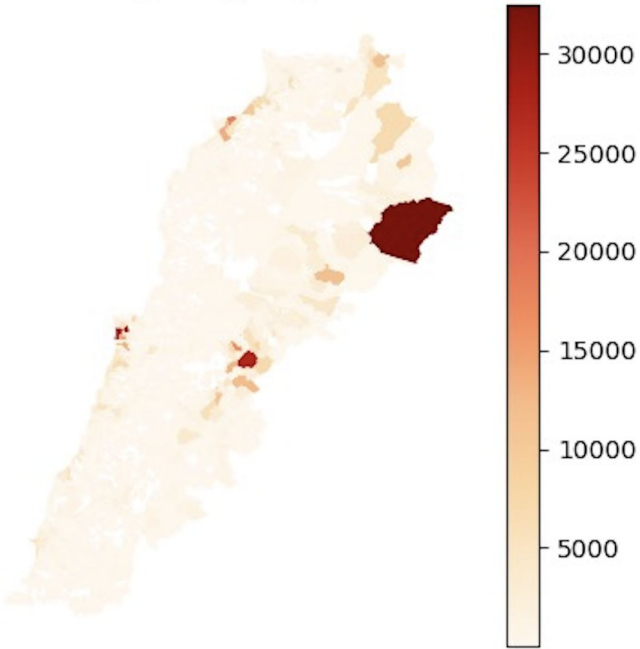
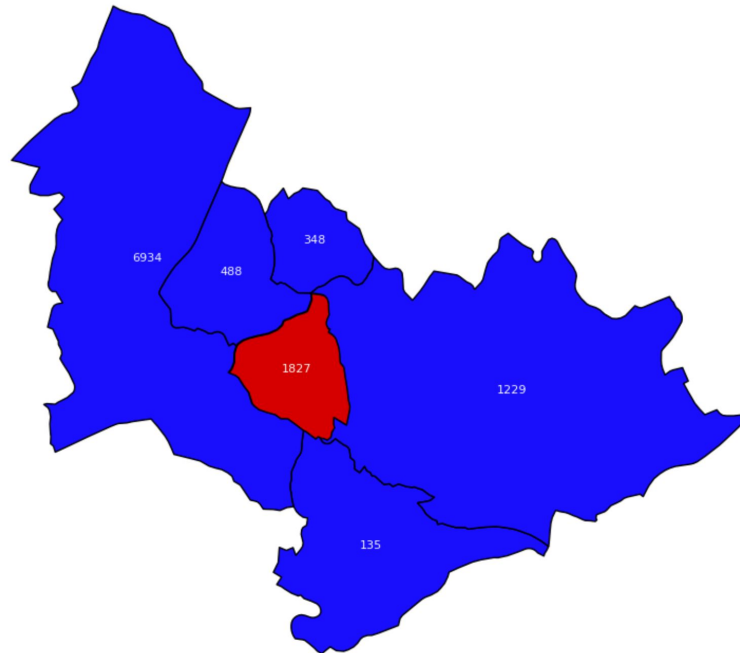# Data Collection Challenges
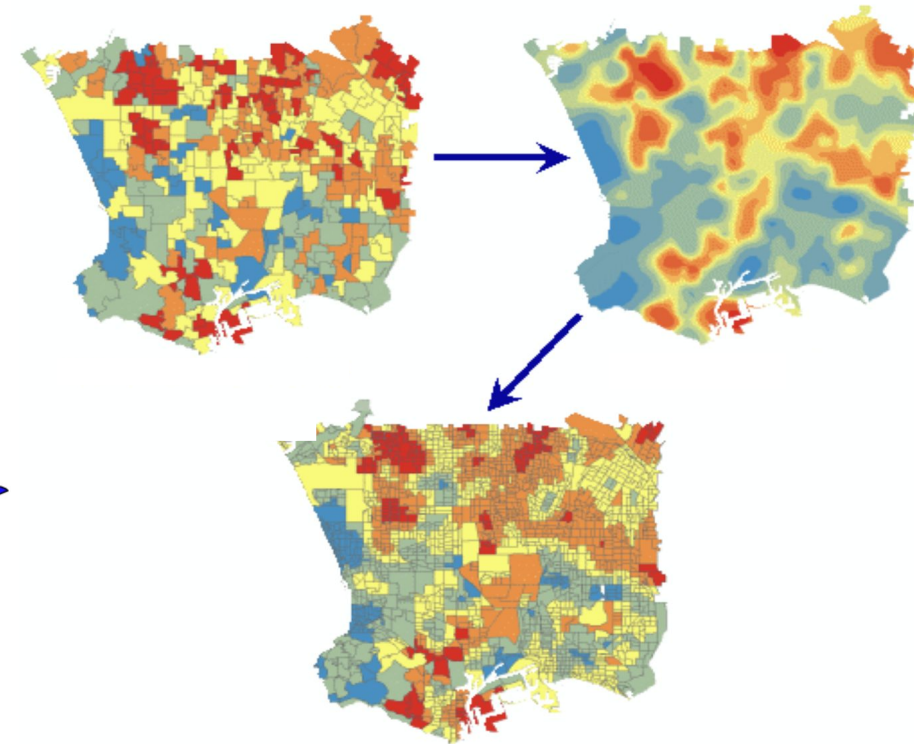
Unavailable Original Datasets
- OCR

Estimated Data
- Interpolation

Inconsistent Data Units
- Disaggregation

Methodology

Preprocess → Model Stable Years →
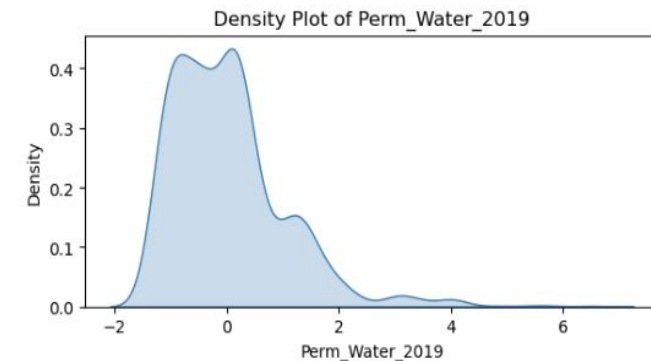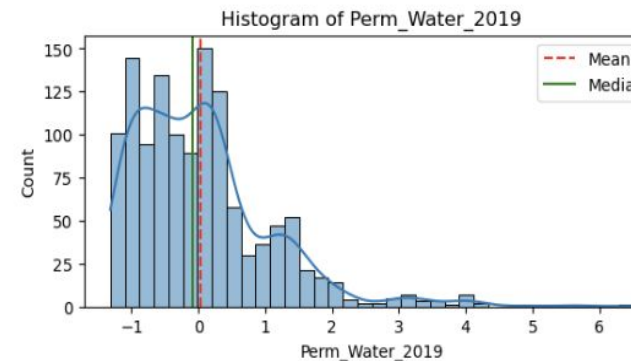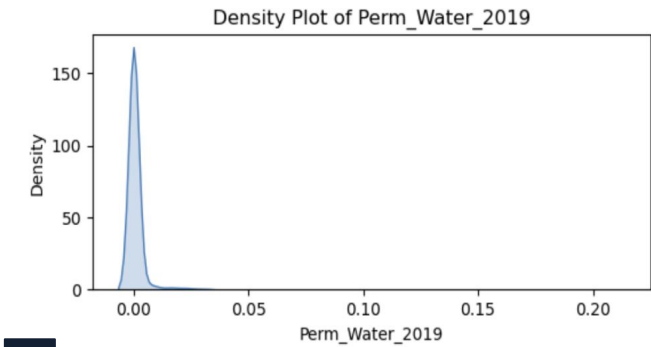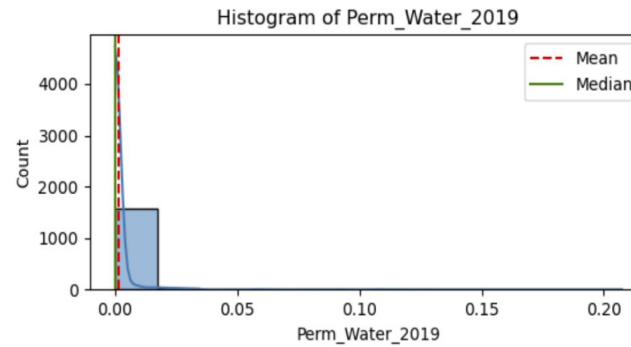Model COVID Years → Conclude

Preprocessing

3 types:

- Proximity Interpolation
- Log
- Standardization

Challenges:

- Zero-inflated features
- Multicollinearity

Models

Models:
- Random Forest
- Multilayer Perceptron (MLP)
- Gradient Boosting
- Support Vector Machine (SVM)
- Linear/Ridge/Lasso Regression
- KNN

Hyperparameter Tuning:
- Grid Search

Metrics:
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-Squared

Model Accuracy Vs Stability

Accuracy Vs Stability:

- Accuracy: how close model predictions are to outcomes ⬅ use various model metrics
- Stability: how robust model is to different sets of data ⬅ look at different years

Our Approach:

1. Run models on 2018, 2019, 2022 (stable years)
2. Run models on 2020, 2021 (COVID years)
3. See if the best model + features importances are similar among both sets of years

Results

Preliminary Results

Best Performing Model: Random Forest
- Lowest MSE/MAE on average
- Economic data had minimal impact

Top Features:
- Population
- Informal Settlements
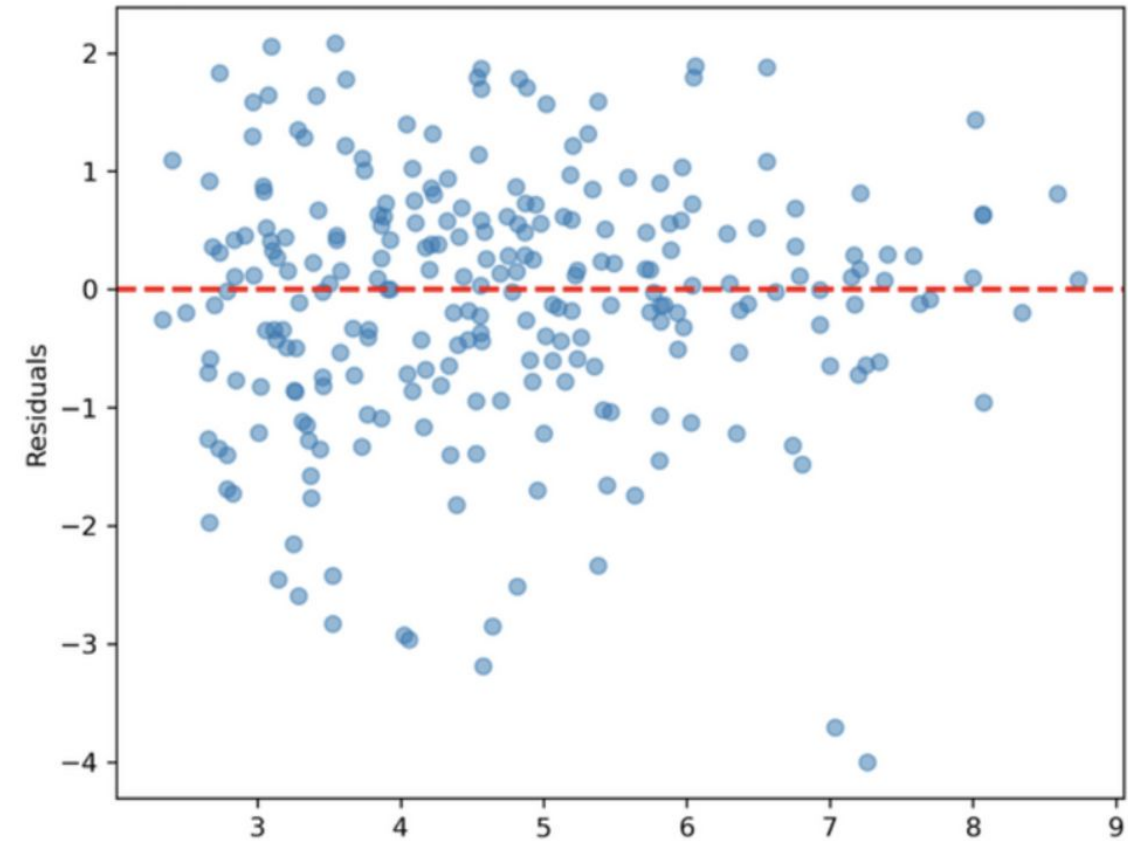- Sunni Religion
- Built Up
- Incoming Roads

Figure: Refugee Num 2022 log Residual Plot by MSE

ILLINOIS NCSA

Next Steps

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

ILLINOIS NCSA

Questions?

ILLINOIS

NCSA | National Center for
Supercomputing Applications