This project provides a tool to determine mitigation potential for towns in floodplains with repetitive losses due to flooding. It uses a calibrated statistical model, specifically a Random Forests based classifier to find potential relocation sites for towns. Model calibration was based on the relocation of 11 Midwestern towns that moved from locations in a floodplain to locations outside of a floodplain between 1937 and 2008.

# Towns for Model Calibration

Model calibration was built around 11 towns in the United States, primarily in the Mid-West, that had relocated after major floods. Other towns that have relocated completely exist in the literature, but they moved for reasons other than flood, so were excluded to not confound the model. The towns included in the model’s calibration dataset are listed in the table below.

|  |  |  |
| --- | --- | --- |
| Town | State | Year Moved |
| Shawneetown | IL | 1937 |
| Leavenworth | IN | 1937 |
| Niobrara | NE | 1969 |
| Odanah | WI | ~1974 |
| Soldiers Grove | WI | 1978 |
| English | IN | 1990 |
| Valmeyer | IL | 1993 |
| Pattonsburg | MO | 1993 |
| Rhineland | MO | 1993 |
| Silex | MO | 1994 |
| Gays Mills | WI | 2008 |

Town boundaries for assessing relocation were based on the location of digitized structures in the town area at the time of the move and then again using recent imagery. Pre-move structures were digitized from the most accurate maps or imagery available before the indicated date. All structures were digitized as point features and assumed to be permanent structures except in the case of Greensburg, KS. Maps were downloaded from the USGS quad map viewer (http://ngmdb.usgs.gov/maps/TopoView/viewer) as georeferenced tiffs. All maps were at a 1:24000 scale if not otherwise noted below. Any imagery used was downloaded from the USGS EarthExplorer (http://earthexplorer.usgs.gov/) datasets as single frame imagery or past NAIP imagery. Historical imagery source and date are listed below. All imagery was georeferenced to the ESRI imagery basemap using a minimum of six geographic control points.

|  |  |  |  |
| --- | --- | --- | --- |
| Town | Type | Source | Date |
| English | USGS Historical Topographic Map | USGS topoView, English, IN | From aerial photographs taken in 1987 and edited in 1989 |
| Gays Mills | USGS Historical Topographic Map | USGS topoView, Gays Mills, WI | Topography from aerial photographs taken 1981, field checked in 1982, with additional map edits in 1983 |
| Leavenworth 1899 | Sanborn-Perris Map Company Limited Surveyers Map | Indiana University Herman B Wells Library Map Collection, https://libraries.indiana.edu/union-list-sanborn-maps#Leavenworth | September 1899 |
| Leavenworth 1927 | Sanborn Map Company Surveyers Map | Indiana University Herman B Wells Library Map Collection, https://libraries.indiana.edu/union-list-sanborn-maps#Leavenworth | May-27 |
| Leavenworth 1947 | USGS Historical Topographic Map | USGS topoView, Leavenworth, IN | 1947 |
| Niobrara | USGS Historical Topographic Map | USGS topoView, Niobrara, NE and Verdigre, NE | Topography from aerial imagery taken 1946 and plane-table surveys from 1950. Field checked in 1950 |
| Odanah | USGS Historical Topographic Map | USGS topoView, Odanah, WI | Topography from aerial photographs taken in 1963 and field checked in 1964 |
| Pattonsburg | USGS Historical Topographic Map | USGS topoView, Pattonsburg, MO and Coffey, MO | From aerial photographs taken in 1981, field checked in 1982, and edited in 1984 |
| Rhineland | USGS Historical Topographic Map | USGS topoView, Gasconade, MO | Topography from aerial photographs taken in 1970 and field checked in 1974 |
| Shawneetown | USGS Historical Topographic Map | USGS topoView, Shawneetown, IL 1:62500 | 1916 |
| Shawneetown (Post-Move) | USGS Historical Topographic Map | USGS topoView, Shawneetown, IL | Aerial photographs taken 1957 and revised in 1959 |
| Silex | USGS Historical Topographic Map | USGS topoView, Silex, MO and Eolia, MO | 1975 map updated with aerial imagery taken in 1982 |
| Soldiers Grove | USGS Historical Topographic Map | USGS topoView, Gays Mills, WI 1:62500 | Topography from aerial photographs taken in 1965 and field checked in 1966 |
| Valmeyer | USGS Historical Topographic Map | USGS topoView, Valmeyer, IL | Topography from aerial photographs taken in 1986. Field checked in 1989 and edited in 1991 |

After digitizing the structures, a boundary was generated to represent the town based on a convex hull (minimum bounding geometry) of the structures, buffered by 100 meters. Where data was available to differentiate, only structures deemed to be permanent, rather than outbuildings, etc, were included for the convex hull. For many towns, manual exclusion of structures located within the towns modern boundary, but outside of the core town was necessary. Where the pre-move boundary and the post-move boundary overlapped, the pre-move area was excluded from the post-move in order to better represent the parts of the town that were new after moving.

# Calibration Data

To find relocation sites, we represented the landscape as a set of hexagons that stood in for parcel information. We were able to obtain real parcel data for some sites, but unable to obtain it for others, so for consistency we used hexagons as a compact, continuous unit that covers each area of interest. The hexagons had a side length of 142 meters, which produces a hexagon with an area that matches the average area of a real parcel in Monroe County, Illinois, where the town of Valmeyer in this analysis is located. The software built the parcels layer to cover the whole 10-mile search radius for new locations around each town.

## Floodplains

Most floodplain data used in the model came from the April 28th, 2016 version of the National Flood Hazard Dataset (NFHL). The model uses the 100 year floodplain data, zones A and AE, as floodplains to avoid. The NFHL is incomplete though, so floodplain data needed to be filled in using hand digitized polygons from the Digital Flood Insurance Rate Maps (DFIRMs) where data were available.

In the case of lands on Native American reservation lands (Odanah and part of the Niobrara search area), DFIRMs were not available, so we generated a representative floodplain polygon using the National Hydrography Dataset Plus Version 2 data. For each segment of river with more than 250 square kilometers of catchment area upstream, the area of the upstream catchment in square kilometers was divided by 10 and the resulting value in meters used to buffer the segment. This buffered area represented the floodplain for these locations. For example, a stream segment with exactly 300 square kilometers upstream would be buffered on each side by 30 meters to obtain a representative floodplain.

## Roads

To represent roads in the model, we digitized major roads from historical USGS topo maps. While a comprehensive roads layer exists for the United States from the US Census Bureau, there is no dataset representing historical roads at the times the towns moved and using the modern layer would skew the model by showing available modern transportation networks where they did not exist previously. For the model, we used main roads since they provide potential existing connections for transportation, but excluded smaller roads that would not support through traffic. Main roads were chosen based on the quadrangle legend and whether the road ran a large distance across the region, so it could serve as a connecting road to the network.

## Other Variables

We hypothesized, based on the choices of the towns in the model, that towns would prefer to stay on the same side of the river they were flooded by when they moved. To handle this, we developed a concept of a barrier river as way to determine whether parcels were on the same side of the river as the town’s original location or on the other side. For towns that straddled their river, this assessment does not distinguish the two. In the case of one town, the barrier line crosses a small ridgeline to connect two small drainages since the main river itself did not bisect the search area - a requirement for our analysis.

Finally, for rivers data, we used NHDPlus flowlines with provided extra attributes representing drainage area attached. For protected areas, we used the Protected Areas Dataset.

# Model Variables

The model uses the following variables:

* Centroid elevation
* Minimum and maximum elevation
* Minimum and maximum slope
* Distance of parcel centroid to original town boundary
* Minimum and maximum distance to floodplain
* Minimum and maximum distance to major road
* Minimum and maximum distance to a protected area
* Minimum and maximum distance to all rivers
* Minimum and maximum distance to rivers with more than 1,000 square Km upstream
* Minimum and maximum distance to rivers with more than 10,000 square Km upstream
* Whether the parcel is on the same side of the river as the original town location
* Whether the parcel was chosen for relocation (response variable)

We tried the following variables but ultimately rejected them for negligible impact on the model:

* Dominant land use in each parcel
* Larger upstream area breakouts for rivers
* Marking whether or not a parcel was protected itself

To obtain the distance variables for each parcel, the software preprocesses each variable involved and creates a Euclidean distance raster for the variable that covers the study area. These are used to extract the minimum and maximum pixel values within each parcel or town boundary with the Zonal Statistics as Table tool in ArcGIS.

# Code Structure

Once the raw data were compiled we built the model. For the model, we used Python 3.4, bundled with ArcGIS Pro for geospatial analysis. The model itself was used scikit-learn, a Python package commonly used for machine learning. To structure the code and provide for potential future web capabilities, the model was built into the Django framework. The full code is openly available at https://github.com/ucd-cws/suitability-analysis/ without the supporting data.

The following are the steps in the processing prior to model construction and calibration:

1. Load data for each region - regions are the software’s concept for the entire search area for each town and serve as a unit for collecting data
2. Preprocess raw data to create parcels and derived distance surfaces
3. Load the relocation town boundaries
4. Extract processed data from region as attributes on parcels, including marking the chosen move location
5. Do the same as the previous step, but for the original town location instead of the parcels, then use this information to make each attribute on the parcels relative to the original value. For example, if the town elevation was at 280 meters and a parcel’s elevation was 305 meters, then the stored elevation value for the parcel would be 25 meters.
6. Merge the attributes for all towns into one data table and load into the statistical model.

Variables were extracted for both the town boundary and for all parcels. To get a relative measure of each value in a parcel, the model subtracted the value for the original location of the town from the parcel’s value. All values were then scaled and centered so that the mean was 0 and the min and max values were -1 and 1.

# Model Structure

To construct the model, we used a Random Forests Classifier in the scikit-learn package for Python. We developed code that handles data loading from parcels, withholding of validation values, randomization, and validation. Our unit of analysis was the parcel and our response variable was whether or not the parcel was chosen for the relocation site of one of the towns. All parcels for all towns were grouped together into one pool for this analysis.

The model code, called ModelRunner, loads all of the parcel records (n=200496) with the preprocessed attributes from the database. It then randomizes the rows before splitting them, with 90% of records (n=180447) used for model training and 10% (n=20049) used for model validation.

# Model Results

Across 100 calibration and validation runs, 20013 validation records were correctly predicted (99.8%) and 36 were incorrectly predicted. Of those 36 records, 32 were underpredictions, or a failure to mark a suitable relocation site as suitable, while 4 were overpredictions, or marking an unsuitable location as suitable. While we consider the 99.8% correct to be good performance for the model, it is important to consider the imbalance of values in the response variable. Only 603 parcels were chosen for relocation of the more than 200000 records in the full dataset, and though Random Forests handles this imbalance, we get poor predictive power from the perspective of locating suitable sites. In the validation dataset an average of 60 of the records were chosen sites, giving us 53% underprediction of relocation sites. The overall performance of the model indicates that the model does well in determining unsuitable locations, but has room for improvement in predicting suitable locations.

The most dominant factors in the model’s classification of a location as suitable or not were the distance to the original town site (almost 25% of predictive power comes from this variable), the minimum elevation difference in a parcel compared with the old town, the distance to protected areas and the distance to rivers with more than 1000 square kilometers upstream. The full list of feature importances is below.

|  |  |
| --- | --- |
| Feature | Importance in model (sums to 1), sorted most to least |
| Centroid Distance to Original Boundary | 0.241 |
| Minimum Elevation | 0.072 |
| Centroid Elevation | 0.057 |
| Minimum Slope | 0.055 |
| Maximum Distance to Protected Areas | 0.048 |
| Max Distance to All Rivers | 0.048 |
| Minimum Distance to Rivers with 1000 sq Km Upstream | 0.047 |
| Maximum Distance to Rivers with 1000 sq Km Upstream | 0.047 |
| Maximum Distance to Floodplain | 0.046 |
| Minimum Distance to Protected Areas | 0.046 |
| Maximum Elevation | 0.045 |
| Maximum Slope | 0.044 |
| Maximum Distance to Major Roads | 0.041 |
| Minimum Distance to Floodplain | 0.041 |
| Minimum Distance to All Rivers | 0.032 |
| Minimum Distance to Major Roads | 0.030 |
| Minimum Distance to Rivers with 10000 sq Km Upstream | 0.024 |
| Maximum Distance to Rivers with 10000 sq Km Upstream | 0.021 |
| Same Side of River | 0.016 |

If we summarize features that measure different version of the same variable, such as minimum distance within a parcel and maximum distance within a parcel, as well as the many river variables into one, we get the following feature importances:

|  |  |
| --- | --- |
| Feature | Importance in model (sums to 1), sorted most to least |
| Centroid Distance to Original Boundary | 0.241 |
| Distance from Rivers | 0.219 |
| Elevation | 0.174 |
| Slope | 0.099 |
| Distance from Protected Areas | 0.093 |
| Distance to Floodplains | 0.087 |
| Distance to Roads | 0.071 |
| Same Side of River | 0.016 |