**IT REVOLUTION IN AGRICULTURE**

**An analytical approach for effectively deploying irrigation technology in New York State.**

**Introduction to Data Analytics (CIS 512)**

**Buffalo State College, Fall 2016**

Submitted to: Dr. John Coles / Dr. Joaquin Carbonara

Report Prepared By: Curtis Robbins

Datasets and R code available at: <https://github.com/CRobbins77/Project>

**TABLE OF CONTENTS**

Executive Summary

Introduction

Project Justification

Project Approach

Summary of Analyst’s Workflow and Results

* + Data Access
  + Data Source
  + Experimental Procedure

Conclusion

Assessment and Validation of another Project

Appendix

**EXECUTIVE SUMMARY**

Throughout the U.S., farmers are non-systematically responding to changing consumer preferences and technological advancements all while dealing with a host of challenges including competition from factory farms, pressure to subdivide property for development and extreme weather conditions. The goals of this project are to strategically deploy advanced irrigation technology to: (1) strengthen small farms (100 cropped acres or less) located in economically depressed counties throughout New York State and (2) improve local economies by increasing production of agricultural commodities to meet the thriving consumer demand for locally grown foods.

To explore opportunities for deploying the latest state of the art irrigation technology in New York State, a series of datasets were analyzed in R to identify a county or counties for investment. Various analytical approaches were used including descriptive statistics, kernel density plots, data discretization and linear regression analysis to develop a robust matrix that allowed for the classification of farms, poverty, wholesale distributors and drought conditions.

Based on this analysis, the deployment of irrigation technology would have the greatest economic impact on **Chemung County,** given its capacity to **increase cropland production, proximity to wholesale distributors, high rates of poverty, average number of small farms** and **extreme drought conditions.**

For more information on the datasets, please see the Appendix.

**INTRODUCTION**

A recent report[[1]](#footnote-1) by the New York State Department of Labor identified a broad-based set of industries as significant based on job count, above-average job growth and above average annual wage. Significant statewide industries included financial activities, educational services, construction and healthcare to name a few. Agriculture, despite having a total estimated economic impact of $37.6 billion in 2011[[2]](#footnote-2), ironically did not make this list. It was considered too small of an industry when compared with the gross state product of nearly 1.5 trillion. “If New York State were an independent nation, it would rank as the 12th or 13th largest economy in the world, depending upon international currency fluctuations”[[3]](#footnote-3). These industries are a reflection of New York City’s dominance on the economy and given our limited investment of $10 million, the agricultural sector has the potential to stretch our dollars furthest by galvanizing economic growth across the entire State.

The potential for economic growth from farming is profound. It is responsible for direct economic activity of “more than $5.4 billion in commodity sales in New York during 2012, an increase of more than 22 percent from 2007”[[4]](#footnote-4). New York is ranked as a national leader for a number of agricultural commodities including milk, yogurt, apples, onions, sweet corn, tomatoes, etc. What makes this industry even more impressive is that they have been able to accomplish this level of growth with minimal technological advancements.

**PROJECT JUSTIFICATION**

Governor Cuomo has made food access a priority of his administration, promoting locally-grown food in traditionally underserved urban communities through a number of initiatives centered on Farmers’ Markets. The idea is those consumers who purchase locally-sourced food will in turn support small “family” farms and the communities they reside in, while expanding access of healthy, nutritious foods to residents.

It has resulted in a vast supply chain that is trying to keep pace with demand at a time when agriculture remains a rather outdated industry. Farmers are systematically responding to changing consumer preferences and technological advancements all while dealing with a host of challenges including competition from factory farms, pressure to subdivide property for development and extreme weather conditions. For example, in 2016, New York State experienced the most extensive severe drought on record. As a result 24 counties across Upstate New York were designated as a natural disaster area by the federal government.

The goal of this project is to strategically deploy advanced technology to strengthen small farms located in economically depressed communities throughout New York State, while increasing production of agricultural commodities, including organically grown produce to meet the strong consumer demand for local foods, e.g. healthy corner stores, farmers markets, etc.

**PROJECT APPROACH**

It all begins with farmers needing better data to generate more tailored solutions and that is where information technology can play a key role. In collaboration with a new ag-tech company called CropX, we have decided to invest the $10 million in state of the art irrigation technology. As recently publicized in WIRED magazine, CropX provides cost-effective technology to assist farmers in determining precisely how much water to use on their crops; it amounts to an “IT revolution in agriculture”.[[5]](#footnote-5)

The project will prioritize investment in a county that meets the following criteria:

1. High rates of poverty
2. Average or high concentration of small farms (100 cropped acres or less)
3. Proximity to wholesale distributors
4. Severe or extreme drought conditions in 2016
5. Capacity to increase cropland production

**Summary of Analyst’s Workflow and Results**

Data Cleaning

* Fill in missing values
* Identify outliers, noisy data – e.g. linear regression of values
* Correct inconsistent data

Data Integration

* Combine data from multiple sources using primary keys, etc.
* Avoid data redundancy

Data Reduction

* Reduce the dataset size (smallest amount of data to accomplish the task)

Data Discretization – “Divide into intervals”

* Interval labels can replace actual data values (low, medium, high) priorities

**KEY FINDINGS / CONCLUSION**

Phase I of the analysis included merging all of the datasets, converting all integers to numeric values, determining the distribution ranges for identified fields (mean and standard deviation) and classifying each based on these calculated ranges. A subset of data was then pulled using the classification results and filters were applied to refine the analysis. Three counties surfaced as being potential candidates for investment, see Figure 1.

*Figure 1*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **FIPS** | **COUNTY** | **FARMED ACRES** | **CROPPED ACRES** | **TOTAL FARMS** | **% POVERTY** | **WHOLESALE DISTRIBUTORS** | **DROUGHT CONDITIONS** |
| 36015 | CHEMUNG | 35,028 | 13,672 | 334 | 18.2 | 1 | EXTREME |
| 36013 | CHAUTAUQUA | 219,780 | 77,698 | 1,187 | 19.3 | 1 | SEVERE |
| 36075 | OSWEGO | 84,223 | 46,298 | 552 | 19.6 | 2 | SEVERE |

Phase II of the analysis included using the same regression analysis that was used previously to fill in the missing fields for cropped acres in the dataset. Only this time, the equation (cropped acres = .4748 \* farmed acres + 1617.97) is used to determine the additional capacity for increased crop production within each of the identified counties.

In comparison to the other 62 counties in New York State, cropped acres for the identified counties can proportionally increase x% without increasing farmed acres.

Chemung (33%) / Chautauqua (36%) / Oswego (-10%)

Based on these findings, the deployment of irrigation technology would have the greatest economic impact on **Chemung County,** given its capacity to **increase cropland production, proximity to wholesale distributors, high rates of poverty, average number of small farms** and **extreme drought conditions.**

**APPENDIX**

**Datasets**

**Dataset 1 - Wholesale Distributor Data by County FIPS (1 file)**

U.S. Census Bureau, American Fact Finder, 2012 Economic Census of the United States

**Description:**

EC1242A1 - Wholesale Trade: Geographic Area Series - Summary Statistics for the U.S., States, Metro Areas, Counties, and Places: 2012

**Data Source:**

<https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmkk>

Select “Table View” from the menu and use the table tools to show hidden rows and columns.

Verify that the following boxes are checked for FIPS state code, FIPS county code, 2012 NAICS code, meaning of 2012 NAICS code and number of establishments.

Research NAICS Codes that are related to merchant wholesalers tied to farmers coops,farmers markets,local wholesalers, distributors, retail grocery stores, etc.

**NAICS SIC CODES:** <http://siccode.com/en/naicscodes/445230/fruit-and-vegetable-markets>

NAICS 42448 - Fresh Fruit and Vegetable Merchant Wholesalers

NAICS 42459 - Other Farm Product Raw Material Merchant Wholesalers

Download .csv file and save as ny\_county\_wholesale

Open file in MS Excel and rename columns accordingly (delete all other fields): County\_FIPS, Name, NAICS\_Code, Description, Year and Total\_Establishments

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Reduce the dataset size to include only NAICS Codes 42448 and 42459**
4. **Rollup codes by county to calculate the total number of wholesale trade establishments by FIPS**
5. **Identify any outliers**

**Dataset 2 - Poverty Data by County FIPS (1 file)**

U.S. Census Bureau - Small Area Income and Poverty Estimates for 2014

**Description:** The U.S. Census Bureau's Small Area Income and Poverty Estimates(SAIPE)program provides annual estimates of income and poverty statistics for all school districts, counties (including those with populations <65K) and states. SAIPE represents the most recent estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions.

**Data Source:** <https://www.census.gov/did/www/saipe/data/statecounty/data/2014.html>

Download .csv file and save as US\_county\_poverty

Open file in MS Excel and rename columns accordingly (delete all other fields): State\_FIPS, County\_FIPS, State, County, Poverty\_Est\_All\_Ages, Poverty\_Per\_All\_Ages, Poverty\_Est\_Age\_0-17, Poverty\_Per\_Age\_ 0-17, Poverty\_Est\_Age\_5-17, Poverty\_Per\_Age\_5-17 and Med\_HH\_Inc

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Reduce the dataset size to include only records for the state of NY**
4. **Remove New York State total records = 62 counties**
5. **Combine State\_FIPS and County\_FIPS to form a five digit FIPS code**
6. **Identify duplicate records or missing values**
7. **Identify any outliers**

**Dataset 3 - Drought Data by County FIPS (2 files)**

USDA – National Drought Mitigation Center, FSA Eligibility Tool: Summary Data (2014 Criteria)

**Description:** Provides county-level data by state to determine, which counties meet the Livestock Forage Disaster Program requirements.

**Data Source:** <http://droughtmonitor.unl.edu/fsa/FsaEligibilityState2014.aspx>

Verify U.S. Drought Monitor Classification Scheme to pull only those counties with severe (D2), extreme (D3) or exceptional drought (D4) conditions.

**DROUGHT CLASS:** <http://droughtmonitor.unl.edu/aboutus/classificationscheme.aspx>

Following the DMC Scheme, two datasets are available (D2) and (D3); (D4) does not apply to any counties in New York State.

**D2 DEFINITION** – Severe drought conditions for at least eight consecutive weeks during the grazing period.

Location: New York State

Grazing Period: Start (01/01/2016) / End (12/31/2016)

Download .csv file and name ny\_county\_D2

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS, State, County, D2\_Class (ConsecWeeks)

**D3 DEFINITION** – Extreme drought conditions for at least four (nonconsecutive) weeks during the grazing period.

Location: New York State

Grazing Period: Start (01/01/2016) / End (12/31/2016)

Download .csv file and name ny\_county\_D3

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS, State, County, D3\_Class (NonConsecWeeks)

**Pre-Processing Steps in R:**

**Both datasets are small enough that all pre-processing steps can be done in MS Excel.**

**Dataset 4 - Farm Data by County (1 file)**

Data.NY.Gov - County Agricultural Districts Profile

**Description:** Includes data on agricultural districts in New York State including towns affected, total acres, farmed acres, cropped acres and number of farms.

**Data Source:** <https://data.ny.gov/Economic-Development/County-Agricultural-Districts-Profile/9bc8-mx4a>

The database is quite robust since the user can filter information prior to exporting it, saving time in pre-processing.

**TIP: Use the Filter feature in the menu to roll-up the pertinent variables (function “sum”) grouped by county.**

FILE ACCESS: Two available options

1. Export file as a standard .csv
2. Access the API – Requires an app token. To sign-up for an app token, click on this **link:** <https://data.ny.gov/login>

It will require you to first set-up an account

**NYSTATE Socrata ID**

**e-mail:** [crobbins@oishei.org](mailto:crobbins@oishei.org)

**password: Nysdata2016**

Next you will need to apply for an app token by clicking on the following link:

<https://dev.socrata.com/foundry/data.ny.gov/8jaw-iviy>

App Token

hpiUm07LyAXAeLZk97XayMxS4

**R CODE:**

**#install.packages("RSocrata")**

**#library("RSocrata")**

**#df <- read.socrata("https://data.ny.gov/resource/8jaw-iviy?$$app\_token= hpiUm07LyAXAeLZk97XayMxS4")**

**#API Endpoint:** <https://data.ny.gov/resource/8jaw-iviy.json>

If option 1, then download .csv file and name ny\_county\_farms

Open file in MSExcel and rename columns accordingly (delete all other fields): County, Towns\_Affected, Total\_Acres, Farmed\_Acres, Cropped\_Acres, Acres\_Owned, Acres\_Rented and Total\_Farms

**TIP: Since the FIPS code is not included in the file, one must use a NY County FIPS Codes lookup table to append the codes for the 62 counties.**

**Data Source:** <https://data.ny.gov/Government-Finance/New-York-State-ZIP-Codes-County-FIPS-Cross-Referen/juva-r6g2/data>

Download .csv file and name ny\_county\_FIPS

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS and County

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Append County FIPS to County Name**
4. **Identify duplicate records or missing values (see below)**

* **Westchester County – Cropped Acres and Acres Rented**
* **Seneca County – Cropped Acres**
* **Putnam County – Cropped Acres and Acres Rented**

1. **Identify any outliers**
2. **Replace missing values using a regression analysis**

1. <http://www.labor.ny.gov/stats/PDFs/Significant-Industries-New-York-State.pdf> [↑](#footnote-ref-1)
2. <http://www.osc.state.ny.us/reports/importance_agriculture_ny.pdf> [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/Economy_of_New_York> [↑](#footnote-ref-3)
4. <http://www.osc.state.ny.us/reports/importance_agriculture_ny.pdf> [↑](#footnote-ref-4)
5. <https://www.wired.com/2015/06/smart-sensor-farmers-dont-waste-water-drought/> [↑](#footnote-ref-5)