CS293S: Internet of Things

An In-Depth Analysis on Weather Data from CIMIS: Estimating Evapotranspiration (ET) Values

Nazmus Saquib Udit Paul Alex Ermakov Santha Ramamoorthy

Graduate Students
Department of Computer Science
University of California Santa Barbara

February 13, 2019

Outline

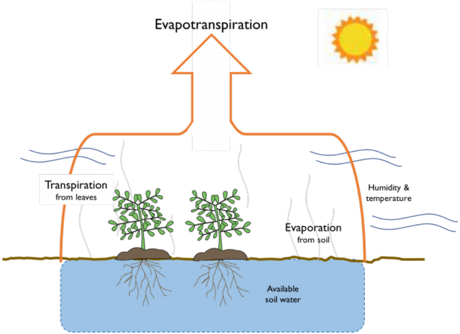
1. Introduction
2. Data Collection
3. Data Overview
4. Regression Analysis
5. Nearest Neighbor Analysis
6. Conclusion
7. Questions

Outline

1. Introduction
2. Data Collection
3. Data Overview
4. Regression Analysis
5. Nearest Neighbor Analysis
6. Conclusion
7. Questions

Introduction: Evapotranspiration (ET)

- Loss of water through:
  1. Evaporation and
  2. Transpiration
- Applications:
  - Irrigation scheduling
  - Water resource planning, etc.



Introduction: CIMIS Weather Stations

- *California Irrigation Management Information System*
- 257 CIMIS stations all through California
    - 136 actively reports ET values
- Measures various weather parameters
- some directly influence ET
- Also measures (*calculates?*) ET

Outline

1. Introduction
2. Data Collection
3. Data Overview
4. Regression Analysis
5. Nearest Neighbor Analysis
6. Conclusion
7. Questions

Data Collection

- Publicly available API
- Reports both hourly and daily data
- A record contains 16 different features
- Current working dataset: data of last one year
- Working dataset will be extended to multiple years:
  - better for capturing seasonal variations

Outline

1. Introduction

2. Data Collection

3. Data Overview

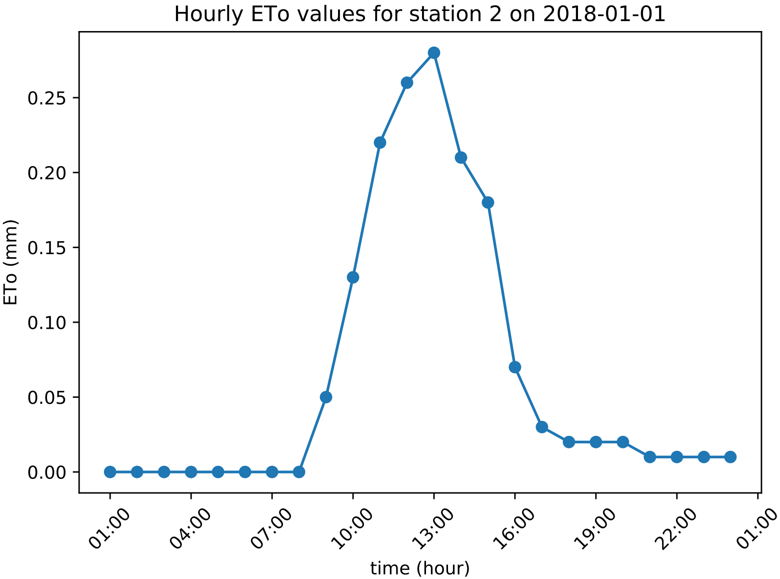
4. Regression Analysis

5. Nearest Neighbor Analysis

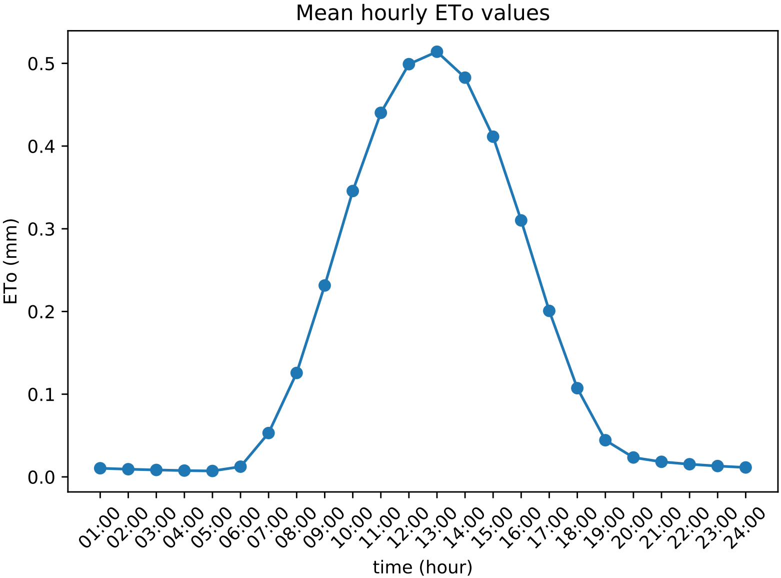
6. Conclusion

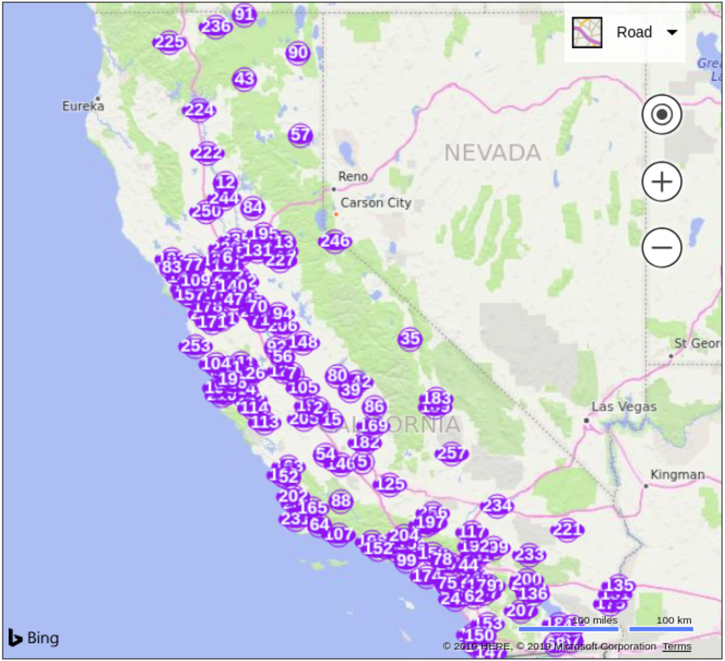
7. Questions

Sample Hourly ET Values



Mean Hourly ET Values

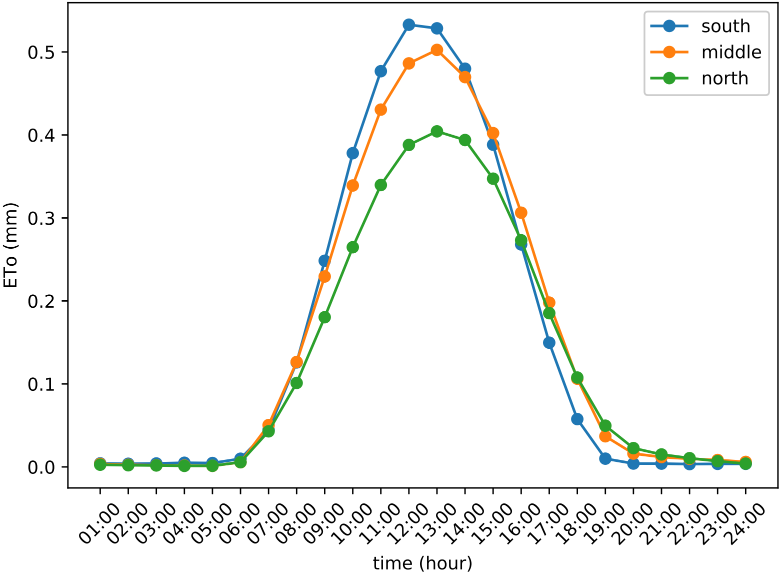



Road
Eureka
NEVADA
Reno
Carson City
CALIFORNIA
St Geor
Las Vegas
Kingman
Gre
La
Bing
100 miles
100 km
© 2018 HERE, © 2019 Microsoft Corporation Terms

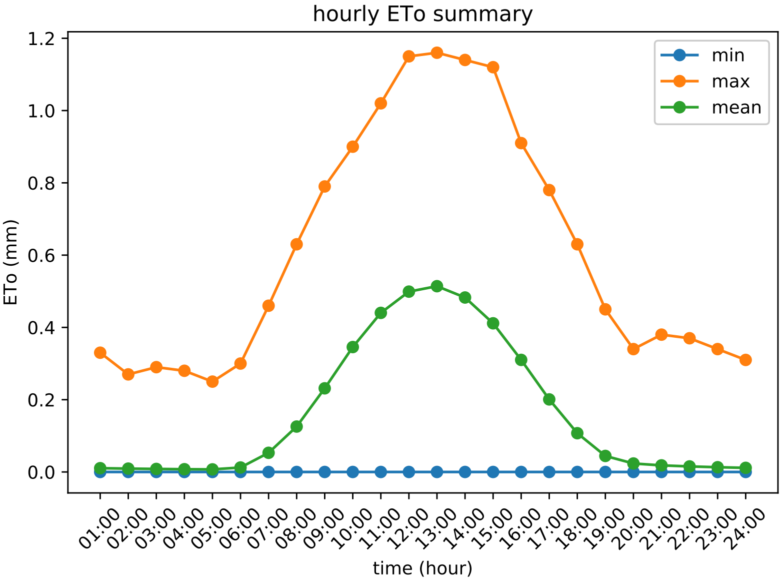
Stations of Interest

- Station with lowest latitude LAT_{MIN} (south)
- Station with highest latitude LAT_{MAX} (north)
- Station with latitude closests to $\frac{LAT_{MIN}+LAT_{MAX}}{2}$ (middle)

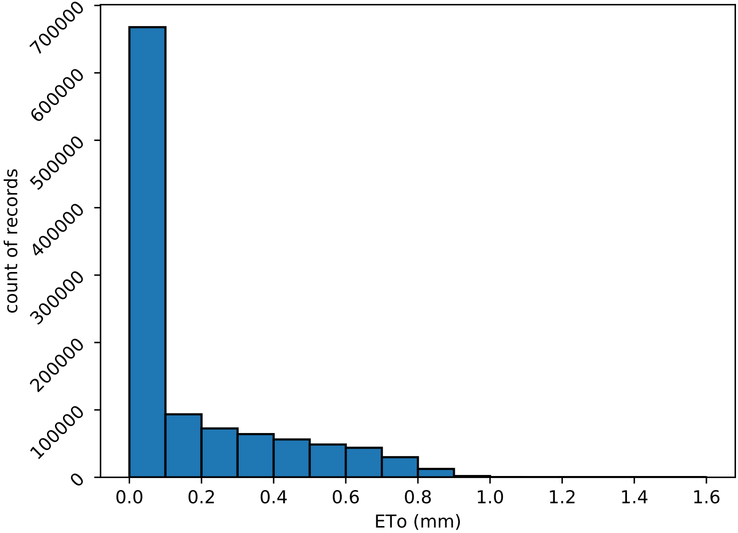
Mean Hourly ET Values of Stations of Interest



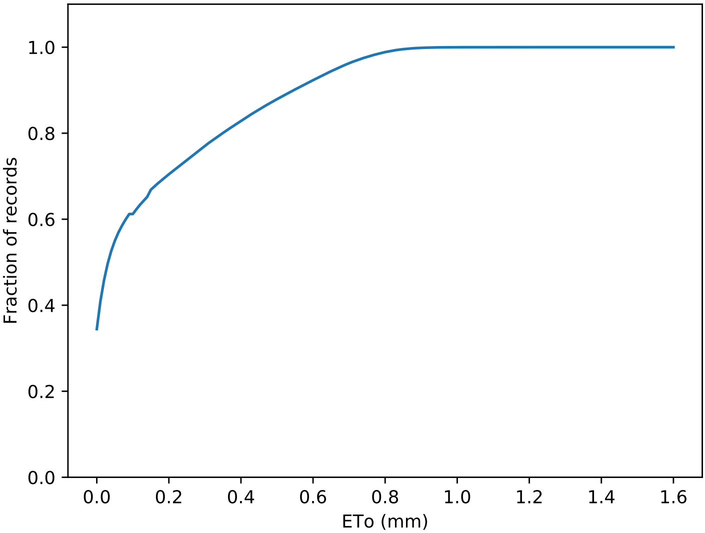
Min/Mean/Max Hourly ET Values



Histogram of ET Values



Empirical CDF of ET Values



Outline

1 Introduction

2 Data Collection

3 Data Overview

4 Regression Analysis

5 Nearest Neighbor Analysis

6 Conclusion

7 Questions

Estimation of ET Values

Given a set of features, can we estimate ET?

- Which features to choose?
- *How well* is our estimate?

Estimation of ET Values

Given a set of features, can we estimate ET?

- Which features to choose?
- *How well* is our estimate?

(CIMIS) Penman Monteith Equation for Calculating ET

$$\boxed{ET_o = \frac{\triangle(R_n - G)}{\lambda[\triangle + \gamma(1 + C_d u_2)]} + \frac{\gamma \frac{37}{T_a + 273.16} u_2 (e_s - e_a)}{\triangle + \gamma(1 + C_d u_2)}}$$

Ultimately depends on four weather features

- Solar net radiation
- Vapor pressure
- Air temperature
- Wind speed

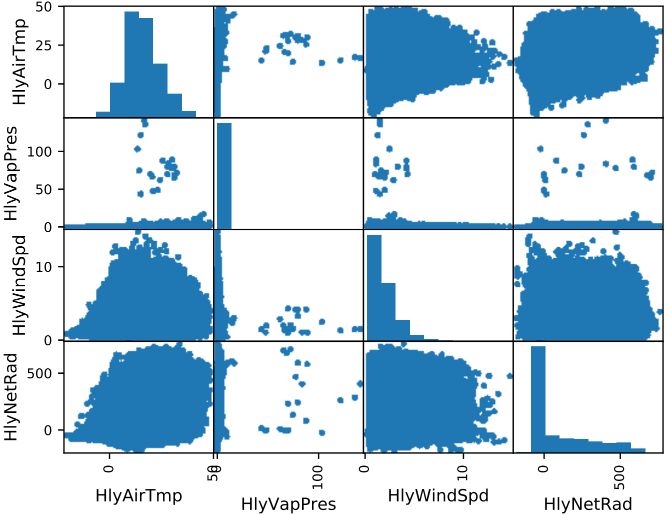
(CIMIS) Penman Monteith Equation for Calculating ET

$$\boxed{ET_o = \frac{\triangle(R_n - G)}{\lambda[\triangle + \gamma(1 + C_d u_2)]} + \frac{\gamma \frac{37}{T_a + 273.16} u_2(e_s - e_a)}{\triangle + \gamma(1 + C_d u_2)}}$$

Ultimately depends on four weather features

- Solar net radiation
- Vapor pressure
- Air temperature
- Wind speed

Scatterplot Matrix of Features of Interest



Regression Results

Features	**Mean Squared Error**	R^2 **Value**
HlyAirTmp,HlyNetRad,HlyVapPres,HlyWindSpd	0.000970123960314	0.9812940161
HlyAirTmp,HlyNetRad,HlyVapPres	0.00130358866256	0.9747612206
HlyAirTmp,HlyNetRad,HlyWindSpd	0.00131186536214	0.9745279825
HlyAirTmp,HlyNetRad	0.00173654973306	0.9665370047
HlyNetRad,HlyVapPres,HlyWindSpd	0.00248645097725	0.9520098573
HlyNetRad,HlyWindSpd	0.0024909080494	0.9516599092
HlyNetRad,HlyVapPres	0.00302176798112	0.9410658003
HlyNetRad	0.00304665078019	0.9409558541
HlyAirTmp,HlyVapPres,HlyWindSpd	0.0236668111725	0.540318481
HlyAirTmp,HlyWindSpd	0.0242823252297	0.5285606181
HlyAirTmp,HlyVapPres	0.026563048828	0.4850281600
HlyAirTmp	0.0278295291341	0.4597101537
HlyVapPres,HlyWindSpd	0.0407552684279	0.2088275258
HlyWindSpd	0.0412914020576	0.1961185540
HlyVapPres	0.0510006461517	0.0128578989

Regression Results

Features	**Mean Squared Error**	R^2 **Value**
HlyAirTmp,HlyNetRad,HlyVapPres,HlyWindSpd	0.000970123960314	0.9812940161
HlyAirTmp,HlyNetRad,HlyVapPres	0.00130358866256	0.9747612206
HlyAirTmp,HlyNetRad,HlyWindSpd	0.00131186536214	0.9745279825
HlyAirTmp,HlyNetRad	0.00173654973306	0.9665370047
HlyNetRad,HlyVapPres,HlyWindSpd	0.00248645097725	0.9520098573
HlyNetRad,HlyWindSpd	0.0024909080494	0.9516599092
HlyNetRad,HlyVapPres	0.00302176798112	0.9410658003
HlyNetRad	0.00304665078019	0.9409558541
HlyAirTmp,HlyVapPres,HlyWindSpd	0.0236668111725	0.540318481
HlyAirTmp,HlyWindSpd	0.0242823252297	0.5285606181
HlyAirTmp,HlyVapPres	0.026563048828	0.4850281600
HlyAirTmp	0.0278295291341	0.4597101537
HlyVapPres,HlyWindSpd	0.0407552684279	0.2088275258
HlyWindSpd	0.0412914020576	0.1961185540
HlyVapPres	0.0510006461517	0.0128578989

Outline

1. Introduction
2. Data Collection
3. Data Overview
4. Regression Analysis
5. Nearest Neighbor Analysis
6. Conclusion
7. Questions

Nearest Neighbor Analysis

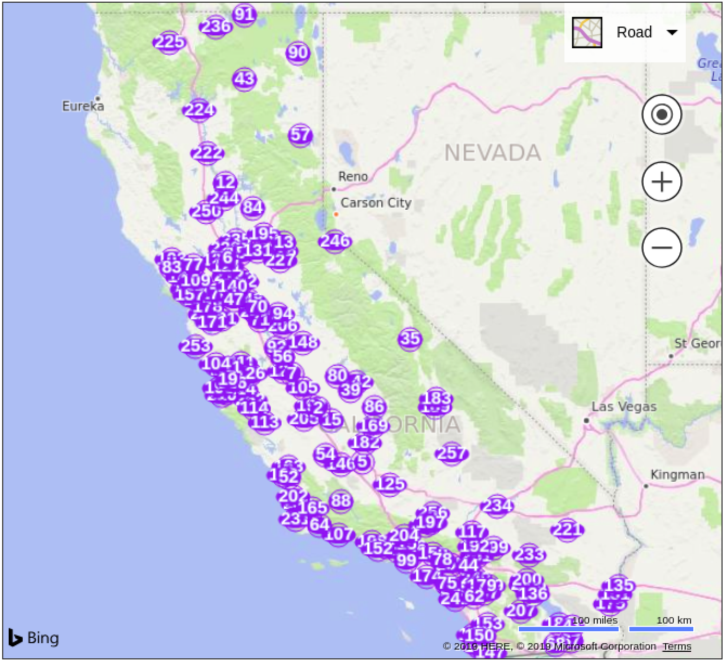
Given the ET value of k nearest stations of a place, can we estimate ET?

- Arithmetic mean of k values
- Inverse Distance Weighted (IDW) average of k values

Nearest Neighbor Analysis

Given the ET value of k nearest stations of a place, can we estimate ET?

- Arithmetic mean of k values
- Inverse Distance Weighted (IDW) average of k values

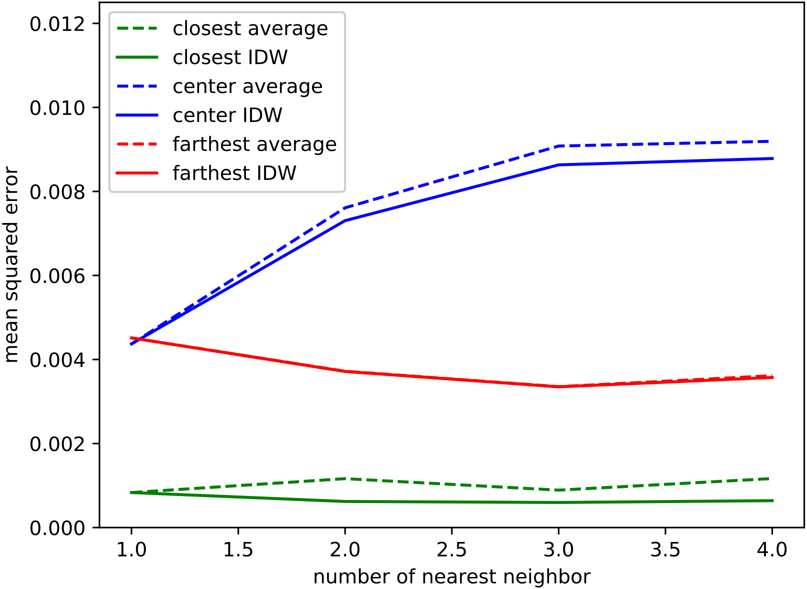

Road
Eureka
Reno
Carson City
NEVADA
CALIFORNIA
Las Vegas
Kingman
St Geor
Gre
91
236
225
90
43
224
57
222
12
244
84
250
246
131
13
227
83
77
6
109
140
157
47
70
94
171
206
253
148
56
104
126
105
80
39
35
183
86
114
113
15
169
182
257
54
146
5
152
125
88
165
64
107
234
197
117
221
204
99
78
44
239
233
200
136
135
207
153
150
Bing
100 miles
100 km
© 2018 HERE, © 2019 Microsoft Corporation Terms

Stations of Interest

- Station with lowest distance D_{MIN} to nearest neighbor
- Station with highest distance D_{MAX} to nearest neighbor
- Station with nearest neighbor at a distance closest to $\frac{D_{MIN}+D_{MAX}}{2}$

Nearest Neighbor Results

Station Number	Num of Neighbors	MSE for Average	MSE for IDW
129	1	0.000832971114168	0.000832971114168
234	1	0.00437018526497	0.00437018526497
57	1	0.00451400872516	0.00451400872516
129	2	0.00116361600992	0.000620877927137
234	2	0.00761026004119	0.00730456269316
57	2	0.00371994564336	0.0037154634375
129	3	0.000890784115612	0.000596760525931
234	3	0.00908058999082	0.00863260116925
57	3	0.00335367604618	0.00334925237208
129	4	0.00116647617403	0.00063999172153
234	4	0.00919325287807	0.00878339044833
57	4	0.00361403432169	0.00357201358681


closest average
closest IDW
center average
center IDW
farthest average
farthest IDW
mean squared error
0.000
0.002
0.004
0.006
0.008
0.010
0.012
1.0
1.5
2.0
2.5
3.0
3.5
4.0
number of nearest neighbor

A Different Approach to Nearest Neighbor

- Some stations are sparsely located, some are densely located
- Distance to *n*th nearest station for different stations might vary widely

What is an optimal value of radius R such that $k\prime$ stations within that radius gives best overall estimates?

work in progress. . .

A Different Approach to Nearest Neighbor

- Some stations are sparsely located, some are densely located
- Distance to nth nearest station for different stations might vary widely

What is an optimal value of radius R such that $k\prime$ stations within that radius gives best overall estimates?

work in progress. . .

Outline

1. Introduction
2. Data Collection
3. Data Overview
4. Regression Analysis
5. Nearest Neighbor Analysis
6. Conclusion
7. Questions

Future Work

- 4 features used in equation, do the other 12 have significant effect on ET?
- Can the dimensionality of dataset be reduced using PCA/LDA?
- Given one (or more, but not all) sensor value at a particular place, how well can we estimate ET by taking into account other sensor values for nearby stations?
- Integrate web interface for analysis.
- Cross check with ground truth value of other sources (*Problem:* features seem to be different?)

Future Work

- 4 features used in equation, do the other 12 have significant effect on ET?
- Can the dimensionality of dataset be reduced using PCA/LDA?
- Given one (or more, but not all) sensor value at a particular place, how well can we estimate ET by taking into account other sensor values for nearby stations?
- Integrate web interface for analysis.
- Cross check with ground truth value of other sources (*Problem:* features seem to be different?)

Future Work

- 4 features used in equation, do the other 12 have significant effect on ET?
- Can the dimensionality of dataset be reduced using PCA/LDA?
- Given one (or more, but not all) sensor value at a particular place, how well can we estimate ET by taking into account other sensor values for nearby stations?
- Integrate web interface for analysis.
- Cross check with ground truth value of other sources (*Problem:* features seem to be different?)

Future Work

- 4 features used in equation, do the other 12 have significant effect on ET?
- Can the dimensionality of dataset be reduced using PCA/LDA?
- Given one (or more, but not all) sensor value at a particular place, how well can we estimate ET by taking into account other sensor values for nearby stations?
- Integrate web interface for analysis.
- Cross check with ground truth value of other sources (*Problem: features seem to be different?*)

Future Work

- 4 features used in equation, do the other 12 have significant effect on ET?
- Can the dimensionality of dataset be reduced using PCA/LDA?
- Given one (or more, but not all) sensor value at a particular place, how well can we estimate ET by taking into account other sensor values for nearby stations?
- Integrate web interface for analysis.
- Cross check with ground truth value of other sources (*Problem:* features seem to be different?)

Outline

1 Introduction

2 Data Collection

3 Data Overview

4 Regression Analysis

5 Nearest Neighbor Analysis

6 Conclusion

7 Questions

Questions?