This presentation premiered at WaterSmart Innovations

watersmartinnovations.com
Dealing with Consumption Data Outliers During Conservation Planning

John M. Clayton, Jack Kiefer, and Lisa Krentz – Hazen and Sawyer
Dave Bracciano, Nisai Wanakule, and Tirusew Asefa – Tampa Bay Water

WaterSmart Innovations Conference and Exposition
October 4th, 2017
South Point Hotel and Conference Center, Las Vegas, NV
Agency Background

Regional water supply authority serving over 2.4 million customers

Six member governments, across three counties

Member demands:
- 2015: 227 MGD
- 2035: 281 MGD (baseline)
Long-Term Demand Forecasting System (LTDFS)

- LTDFS designed to:
  - Track water consumption, socioeconomic, economic and policy conditions
  - Provide inputs for demand forecasting models (updated periodically)
  - Prepare forecasts through implementation of models (annually)
  - Inform regional and member specific demand management efforts
  - Support water supply reliability ("just-in-time" supply development) efforts
Database objectives

Extensive LTDFS database effort to:

1. Provide water use data and property characteristics for all individual customers (locations) with ability to aggregate to larger geographies
2. Ensure acquired information can be maintained through time to support future evaluations
3. Standardize design so queries and analytical routines can be replicated and updated efficiently through time
Information developed for each location

• Water use class
  • Retail/billing
  • Property use code
• Historical sales of potable water
  • Monthly (1998-2016)
  • Domestic meter(s)
  • Irrigation meter(s)
• Access to reclaimed water

• Property characteristics
  • Dwelling units (residential)
  • Year built
  • Lot size/area
  • Building/heated area
  • Other
Locations and Small-Scale Geographies = More Noise in Consumption Data!

Hundreds of thousands of locations, tens of millions of monthly consumption points

Outliers can be anywhere

Potential to obfuscate or bias small-scale analyses

Can we manually spot and correct/flag them all?
Locations and Small-Scale Geographies = More Noise in Consumption Data!

Hundreds of thousands of locations, tens of millions of monthly consumption points
Outliers can be anywhere
Potential to obfuscate or bias small-scale analyses
Can we manually spot and correct/flag them all?

**NO WAY!**

We need automated screening procedures
SF Consumption in Tampa Bay

What is typical and what is not?

Single-family non-irrigator in Tampa area: 100-200 gal/day (gpd) average across a month

One irrigation cycle might dispense 2500 gallons

- 1 irrigation/week: 450-550 gpud in a month
- 2 irrigations/week: 750-850 gpud in a month
- 3 irrigations/week: 1100-1200 gpud in a month
- Daily irrigation: 2600-2700 gpud in a month
What Is An Outlier?

Physically speaking...

SF HH consumption becomes more physically unreasonable as it increases beyond about 2000 gpwd

Leaks? Billing corrections/irregularities not related to actual use?
What Is An Outlier?

Also depends on how individual records relate to overall trend, seasonality at each SF household.

Both can change over time (changing customers at same household, changing fixtures and efficiency)

Outliers

Decreasing overall trend and smaller seasonality… may not be so OK out here
What Is An Outlier?

Outliers can be physically reasonable but way out of character for a given household.
What Is An Outlier?

Sometimes outlier status is not obvious

Somewhat out of character, but these are in Spring (hot/dry) season
Also, previous year had high Spring consumption
Several Common Screening Methods

Global gpd threshold

One threshold does not fit all

Individual gpd thresholds (e.g. top n% for each household)

Not all households really have outliers

Strong seasonality and changing patterns over time – could discard real and critical data for our analyses

Neither approach has literature-based statistical guidance on outlier detection
New Screening Method for Tampa Bay Water

1) Bulk-screen monthly SF consumption records
   → Peak gpd > some physically-based threshold

2) Detrend and deseasonalize monthly gpd series for each household
   → Provides series of normalized residuals

3) Analyze residual time series to detect outliers
   → Data points that stand out in their own time environment, even after accounting for trend and seasonality
   → Statistical method for normalized data: Cook’s D
# 1) SF Monthly Consumption Screening

Many Options

<table>
<thead>
<tr>
<th>peak gpd threshold</th>
<th>Total Households</th>
<th>% of all Households</th>
<th>total Household/months</th>
<th>% of all Household/months</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>30893</td>
<td>6.3%</td>
<td>3476231</td>
<td>5.9%</td>
</tr>
<tr>
<td>2500</td>
<td>17936</td>
<td>3.6%</td>
<td>2003792</td>
<td>3.4%</td>
</tr>
<tr>
<td>3000</td>
<td>11143</td>
<td>2.3%</td>
<td>1237017</td>
<td>2.1%</td>
</tr>
<tr>
<td>4000</td>
<td>5171</td>
<td>1.0%</td>
<td>571162</td>
<td>1.0%</td>
</tr>
<tr>
<td>5000</td>
<td>2819</td>
<td>0.6%</td>
<td>310777</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

| total SF Households | 492823 |
| total SF Household/months | 59173132 |
2a) Detrend Each Household’s Gpd

Calculate trend

13-month centered weighted moving average of gpd (1/24 on months 1 and 13, 1/12 on months 2-12)
2a) Detrend Each Household’s Gpd

Calculate trend

13-month centered weighted moving average of gpd (1/24 on months 1 and 13, 1/12 on months 2-12)

\[
\text{Detrended} = \frac{\text{gpd}}{\text{trend}}
\]
2b) Deseasonalize Each Household’s Gpd

Calculate seasonal pattern

Regress detrended series using monthly fixed effects
2b) Deseasonalize Each Household’s Gpd

Calculate seasonal pattern

Regress detrended series using monthly fixed effects

Residual = detrended / seasonal
3) Analyze residual time series to detect outliers

Linear regression method to detect outliers: Cook’s D statistic

Common output from regression packages
A statistical value for each data point produced by fitting
Indicates how much each data point influences regression coefficients
Common guidance: if Cook’s D > 4/n then point is outlier

\[
D_i = \frac{(y_i - \hat{y}_i)^2}{(p + 1)s^2 \left[ \frac{h_i}{(1 - h_i)^2} \right]}
\]

- \(D_i\) = Cook’s distance measure for observation \(i\)
- \(y_i - \hat{y}_i\) = the residual for observation \(i\)
- \(h_i\) = the leverage for observation \(i\)
- \(p\) = the number of independent variables
- \(s\) = the standard error of the estimate
3) Analyze residual time series to detect outliers

Fit a regression: $gpd = \text{constant}$
SF Screening Method w/ Cook’s Distance

Many options now…

- Gpd > Fixed threshold (2000, 2500, 3000…)
- Cook’s D > threshold (4/# obs, 8/# obs…)
- Gpd > threshold AND Cook’s D > threshold
- Gpd > threshold OR Cook’s D > threshold
Example

Combining GPD and Cook’s D Thresholds
Example

Combining GPD and Cook’s D Thresholds

Consumption > 2000 gpd: flag these observations
Example

Combining GPD and Cook’s D Thresholds

Consumption > 2000 gpd and D > 4/n: flag these observations
Example

Combining GPD and Cook’s D Thresholds

- Consumption > 2500 gpd and D > 4/n: flag these observations
Example

Combining GPD and Cook’s D Thresholds

Consumption > 3000 gpd and D > 4/n: flag these observations
Example

Combining GPD and Cook’s D Thresholds

Consumption > 4000 gpd and D > 4/n: flag these observations
All Methods Require Judgement

Although outlier detection methods are quantitative, still requires qualitative decisions

Threshold selections

What do about identified outliers
Conclusion

Detrend + Deseasonalize + Analyze Residuals using Cook’s D statistic

Better confidence as an automated method for mass-screening of outliers

- Provides more info for outlier judgement than gpd thresholds: Intel on time-environment of consumption data
- Statistical characterization of departures by established means
- Individual visual assessments still possible
- In absence of visual assessment, analyst still knows there is a rational mechanical basis for identifying the outliers
Thank You!

watersmart INNOVATIONS 2017
# Locations and Observations with Flags

## Cooks4 vs Cooks8

<table>
<thead>
<tr>
<th>gpd threshold</th>
<th>Qualifying Locations</th>
<th>Cook4 1+ month &gt; Cook4 &amp; &gt;gpd thresh</th>
<th>Locations</th>
<th>%</th>
<th>Locations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>32725</td>
<td>32457</td>
<td>99.2%</td>
<td></td>
<td>30314</td>
<td>92.6%</td>
</tr>
<tr>
<td>2500</td>
<td>21992</td>
<td>21790</td>
<td>99.1%</td>
<td></td>
<td>20758</td>
<td>94.4%</td>
</tr>
<tr>
<td>3000</td>
<td>17927</td>
<td>17758</td>
<td>99.1%</td>
<td></td>
<td>16685</td>
<td>93.1%</td>
</tr>
<tr>
<td>4000</td>
<td>15241</td>
<td>15089</td>
<td>99.0%</td>
<td></td>
<td>13588</td>
<td>89.2%</td>
</tr>
<tr>
<td>5000</td>
<td>12024</td>
<td>11903</td>
<td>99.0%</td>
<td></td>
<td>10680</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gpd threshold</th>
<th>Obs in Qualifying Locations</th>
<th>Cook4 months &gt; Cook4 &amp; &gt;gpd thresh</th>
<th>Obs</th>
<th>%</th>
<th>Obs</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>3648048</td>
<td>142269</td>
<td>3.9%</td>
<td></td>
<td>55468</td>
<td>1.5%</td>
</tr>
<tr>
<td>2500</td>
<td>2438028</td>
<td>93843</td>
<td>3.8%</td>
<td></td>
<td>41582</td>
<td>1.7%</td>
</tr>
<tr>
<td>3000</td>
<td>1991471</td>
<td>77491</td>
<td>3.9%</td>
<td></td>
<td>34755</td>
<td>1.7%</td>
</tr>
<tr>
<td>4000</td>
<td>1701716</td>
<td>68380</td>
<td>4.0%</td>
<td></td>
<td>27564</td>
<td>1.6%</td>
</tr>
<tr>
<td>5000</td>
<td>1335844</td>
<td>53452</td>
<td>4.0%</td>
<td></td>
<td>19350</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

| tot SF locations in WYs 2003-2013 | 555019 |
| tot location/months in WYs 2003-2013 | 64945036 |
Frequency of flags per location

2000 gpd threshold