Data Analytics with MATLAB

Adam Filion
Application Engineer
MathWorks
Case Study: Day-Ahead Load Forecasting

- **Goal:**
  - Implement a tool for *easy* and *accurate* computation of day-ahead system load forecast

- **Requirements:**
  - Acquire and clean data from multiple sources
  - Accurate predictive model
  - Easily deploy to production environment
Challenges with Data Analytics

- Aggregating data from multiple sources
- Cleaning data
- Choosing a model
- Moving to production
NYISO Energy Load Data

mis.nyiso.com/public/

Real-Time Actual Load

<table>
<thead>
<tr>
<th>CSV Files</th>
<th>Last Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-21-2014</td>
<td>10/21/14 23:02 EDT</td>
</tr>
<tr>
<td>10-20-2014</td>
<td>10/21/14 00:07 EDT</td>
</tr>
<tr>
<td>10-19-2014</td>
<td>10/20/14 00:01 EDT</td>
</tr>
<tr>
<td>10-18-2014</td>
<td>10/18/14 23:59 EDT</td>
</tr>
<tr>
<td>10-17-2014</td>
<td>10/18/14 00:00 EDT</td>
</tr>
<tr>
<td>10-16-2014</td>
<td>10/16/14 23:59 EDT</td>
</tr>
<tr>
<td>10-15-2014</td>
<td>10/15/14 23:59 EDT</td>
</tr>
<tr>
<td>10-14-2014</td>
<td>10/14/14 23:59 EDT</td>
</tr>
<tr>
<td>10-13-2014</td>
<td>10/13/14 23:59 EDT</td>
</tr>
<tr>
<td>10-12-2014</td>
<td>10/12/14 23:59 EDT</td>
</tr>
</tbody>
</table>

Archived Files (zip format)

<table>
<thead>
<tr>
<th>CSV Files</th>
<th>Last Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-2014</td>
<td>10/21/14 23:02 EDT</td>
</tr>
<tr>
<td>09-2014</td>
<td>09/30/14 23:59 EDT</td>
</tr>
<tr>
<td>08-2014</td>
<td>09/01/14 00:01 EDT</td>
</tr>
<tr>
<td>07-2014</td>
<td>08/01/14 00:00 EDT</td>
</tr>
<tr>
<td>06-2014</td>
<td>07/01/14 00:00 EDT</td>
</tr>
<tr>
<td>05-2014</td>
<td>06/01/14 00:00 EDT</td>
</tr>
<tr>
<td>04-2014</td>
<td>04/30/14 23:59 EDT</td>
</tr>
<tr>
<td>03-2014</td>
<td>03/31/14 23:59 EDT</td>
</tr>
<tr>
<td>02-2014</td>
<td>02/28/14 23:58 EST</td>
</tr>
<tr>
<td>01-2014</td>
<td>01/31/14 23:59 EST</td>
</tr>
</tbody>
</table>
Techniques to Handle Missing Data

- **List-wise deletion**
  - Unbiased estimates
  - Reduces sample size

- **Implementation options**
  - Built in to many MATLAB functions
  - Manual filtering

![Table with data](image-url)
Techniques to Handle Missing Data

Substitution – replace missing data points with a reasonable approximation

Table

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01-Jan-2004 00:00:00</td>
<td>1015</td>
<td>1651</td>
<td>618</td>
<td>972</td>
<td>1120</td>
</tr>
<tr>
<td>2</td>
<td>01-Jan-2004 01:00:00</td>
<td>927</td>
<td>1562</td>
<td>568</td>
<td>905</td>
<td>1019</td>
</tr>
<tr>
<td>3</td>
<td>01-Jan-2004 02:00:00</td>
<td>891</td>
<td>1507</td>
<td>541</td>
<td>858</td>
<td>977</td>
</tr>
<tr>
<td>4</td>
<td>01-Jan-2004 03:00:00</td>
<td>NaN</td>
<td>1440</td>
<td>517</td>
<td>821</td>
<td>927</td>
</tr>
<tr>
<td>5</td>
<td>01-Jan-2004 04:00:00</td>
<td>NaN</td>
<td>1434</td>
<td>499</td>
<td>798</td>
<td>905</td>
</tr>
<tr>
<td>6</td>
<td>01-Jan-2004 05:00:00</td>
<td>NaN</td>
<td>1449</td>
<td>496</td>
<td>805</td>
<td>915</td>
</tr>
<tr>
<td>7</td>
<td>01-Jan-2004 06:00:00</td>
<td>NaN</td>
<td>1490</td>
<td>524</td>
<td>832</td>
<td>931</td>
</tr>
<tr>
<td>8</td>
<td>01-Jan-2004 07:00:00</td>
<td>NaN</td>
<td>1525</td>
<td>526</td>
<td>861</td>
<td>940</td>
</tr>
<tr>
<td>9</td>
<td>01-Jan-2004 08:00:00</td>
<td>960</td>
<td>1529</td>
<td>518</td>
<td>878</td>
<td>962</td>
</tr>
<tr>
<td>10</td>
<td>01-Jan-2004 09:00:00</td>
<td>1046</td>
<td>1628</td>
<td>541</td>
<td>911</td>
<td>1027</td>
</tr>
<tr>
<td>11</td>
<td>01-Jan-2004 10:00:00</td>
<td>1111</td>
<td>1706</td>
<td>570</td>
<td>992</td>
<td>1079</td>
</tr>
</tbody>
</table>

Easy to model

Too important to exclude
Merge Different Sets of Data

- Join along a common axis

- Popular Joins:
  - Inner
  - Full Outer
  - Left Outer
  - Right Outer
## Full Outer Join

### First Data Set

<table>
<thead>
<tr>
<th>Key</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>4</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>1.7</td>
</tr>
<tr>
<td>9</td>
<td>1.9</td>
</tr>
</tbody>
</table>

### Second Data Set

<table>
<thead>
<tr>
<th>Key</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

### Joined Data Set

<table>
<thead>
<tr>
<th>Key</th>
<th>B</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>NaN</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>1.4</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>5</td>
<td>NaN</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>1.7</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>1.9</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>
Learn More: Big Data with MATLAB

www.mathworks.com/discovery/matlab-mapreduce-hadoop.html

MapReduce on the Desktop

Explore and analyze big data sets on your desktop with the MapReduce programming technique built into MATLAB.

Creating algorithms using MapReduce: max, mean, mean by group, histograms, covariance and related quantities, summary statistics by group, logistic regression, tall skinny QR

- Get started with MATLAB MapReduce
- MapReduce design patterns
- Use MATLAB MapReduce with relational databases

MapReduce on Hadoop

Execute MATLAB MapReduce based algorithms within Hadoop MapReduce to explore and analyze data that is stored and managed on Hadoop, using MATLAB Distributed Computing Server.

- Run MATLAB MapReduce on Hadoop

Create applications and libraries based upon MATLAB MapReduce for deployment within production instances of Hadoop, using MATLAB Compiler.

- Deploy MATLAB MapReduce applications to Hadoop

MATLAB® has numerous capabilities for exploring and analyzing big data sets. Among them is MapReduce, a powerful and established programming technique for applying filtering, statistics and other general analysis methods to big data.

The MapReduce functionality built into MATLAB lets you analyze data that does not fit into memory. By running your MapReduce based algorithms in parallel (using Parallel Computing Toolbox™), you can better utilize the processing resources on your desktop without changing your algorithms.

To analyze data in MATLAB using MapReduce:
1. Specify the data you want to analyze using datastore
2. Create your map and reduce functions in MATLAB
3. Execute your map and reduce functions using mapreduce

While MATLAB MapReduce is optimized for array-based analysis, it is fully compatible with Hadoop MapReduce, so you can run your MapReduce based algorithms within the Hadoop MapReduce framework:
- Execute MapReduce based algorithms on Hadoop directly from the MATLAB desktop, using MATLAB Distributed Computing Server™
- Package MapReduce based algorithms for deploying to production Hadoop systems, using MATLAB Compiler™
Challenges with Data Analytics

- Aggregating data from multiple sources
- Cleaning data
  - Choosing a model
  - Moving to production
Machine Learning
Characteristics and Examples

- Characteristics
  - Lots of variables
  - System too complex to know the governing equation (e.g., black-box modeling)

- Examples
  - Pattern recognition (*speech, images*)
  - Financial algorithms (*credit scoring, algo trading*)
  - Energy forecasting (*load, price*)
  - Biology (*tumor detection, drug discovery*)
Overview – Machine Learning

Type of Learning

- Supervised Learning
  - Develop predictive model based on both input and output data
- Unsupervised Learning
  - Group and interpret data based only on input data

Categories of Algorithms

- Classification
- Regression
- Clustering
Supervised Learning

Regression

- Neural Networks
- Decision Trees
- Ensemble Methods
- Non-linear Reg. (GLM, Logistic)
- Linear Regression

Classification

- Support Vector Machines
- Discriminant Analysis
- Naive Bayes
- Nearest Neighbor
Unsupervised Learning

Clustering

- k-Means, Fuzzy C-Means
- Hierarchical
- Neural Networks
- Gaussian Mixture
- Hidden Markov Model
Learn More: Machine Learning with MATLAB

mathworks.com/machine-learning
Challenges with Data Analytics

- Aggregating data from multiple sources
- Cleaning data
- Choosing a model
  - Moving to production
Deployment Highlights

- Share with others who may not have MATLAB
- Royalty-free deployment
- Encryption to protect your intellectual property
Deploying Applications with MATLAB

MATLAB Compiler for sharing MATLAB programs without integration programming

MATLAB Compiler SDK provides implementation and platform flexibility for software developers

MATLAB Production Server provides the most efficient development path for secure and scalable web and enterprise applications
Sharing Standalone Applications

1. Application Author
   - MATLAB Toolboxes
   - MATLAB Application

2. MATLAB Compiler
   - Standalone Application
   - Excel Add-in
   - Hadoop

3. End User
   - MATLAB Runtime
Integrating MATLAB-based Components

Application author and software developer *might* be the same person.
MATLAB Production Server

- Directly deploy MATLAB analytic programs into production
  - Centrally manage multiple MATLAB programs & MCR versions
  - Automatically deploy updates without server restarts

- Scalable & reliable
  - Service large numbers of concurrent requests
  - Add capacity or redundancy with additional servers

- Use with web, database & application servers
  - Lightweight client library isolates MATLAB processing
  - Access MATLAB programs using native data types
  - Integrates with Java, .NET, C and Python
Learn More: Application Deployment with MATLAB

www.mathworks.com/solutions/desktop-web-deployment/

Deploying MATLAB Code as an Executable or Software Component

Using MathWorks application deployment products, you can eliminate the costly and error-prone work of recoding your MATLAB algorithms in another programming language. Because you maintain your source code in MATLAB, you can easily develop and update your algorithms and automatically package them as standalone executables or software components for integration in environments such as C, C++, Java™, .NET, and Excel®.

MATLAB Compiler packages your MATLAB applications as encrypted standalone executables or C/C++ shared libraries. MATLAB builder products work in conjunction with MATLAB Compiler to create standard components for use with Java, .NET, or Excel. These executables and components can be deployed royalty-free on operating systems supported by MATLAB.

Deploying MATLAB Code as a Web Application

Using MathWorks application deployment products, you can develop MATLAB based components for the Web that execute mathematical computations and generate interactive graphics. After developing an algorithm in MATLAB, you can automatically create a standard component designed to integrate in a Web application using MATLAB builder products for either Java or .NET.

Once you place the component on a Web server, your users access the application through a Web browser and do not need to install additional software on their desktop computers.

The Java and .NET components created by the deployment tools can be used in conjunction with standard Web technologies such as ASP.NET, SOA, SaaS, JavaScript, and HTML.
Learn More: MATLAB Application Deployment

Also … www.mathworks.com/solutions/desktop-web-deployment/
Data Analytics Products

Access and Explore Data
Preprocess Data
Develop Predictive Models
Integrate Analytics with Systems

MATLAB

Parallel Computing Toolbox, MATLAB Distributed Computing Server
Database Toolbox
Data Acquisition Toolbox
Mapping Toolbox
Image Acquisition Toolbox
OPC Toolbox

Statistics and Machine Learning Toolbox
Curve Fitting Toolbox
Signal Processing Toolbox
Image Processing Toolbox

MATLAB Production Server
MATLAB Compiler
MATLAB Compiler SDK

Econometrics Toolbox
Computer Vision System Toolbox

Used in today’s demo
Additional Data Analytics products
Key Takeaways

- Data preparation can be a big job; leverage built-in MATLAB tools and spend more time on the analysis.

- Rapidly iterate through different predictive models, and find the one that’s best for your application.

- Leverage parallel computing to scale-up your analysis to large datasets.

- Eliminate the need to recode by deploying your MATLAB algorithms into production.