NOVEC Data Analysis and Forecasting Cost Model

Grant Huang

Craig McClellan

Sajjad Taghiyeh

Junchao Zhuang

Agenda

- Background
- Problem Statement/Description
- Data Analysis
- Cost Forecasting Models
  - Regression
  - Simulation
- Recommendations
- Way Forward

Background

- NOVEC is a non-profit cooperative providing electricity to the counties of Clarke, Fairfax, Fauquier, Loudoun, Prince William, and Stafford
- It has a service territory of 651 square miles with more than 6,880 miles of power lines
- NOVEC serves more than 155,000 residences and businesses in Northern Virginia

Background Cont.

- NOVEC provides electricity to many different kinds of customers:
  - Non-Residential Commercial Customers
    - Businesses
    - Government Offices / Buildings
  - Residential Customers
    - Single Family Homes
    - Town Houses
    - Condominiums
    - Apartment Buildings
    - May include the small substations and other infrastructure needed to supply power to these customers

Problem Statement

- NOVEC tasked us with analyzing a historical data set of construction project costs in an attempt to create a model that can estimate total cost to connect new residential customers in the next few years
- The input will be NOVEC's prediction of new residential customers
- The total cost output should include:
  - Residential construction costs
  - Related ancillary costs associated with Mainline, Infrastructure, and "Other" construction

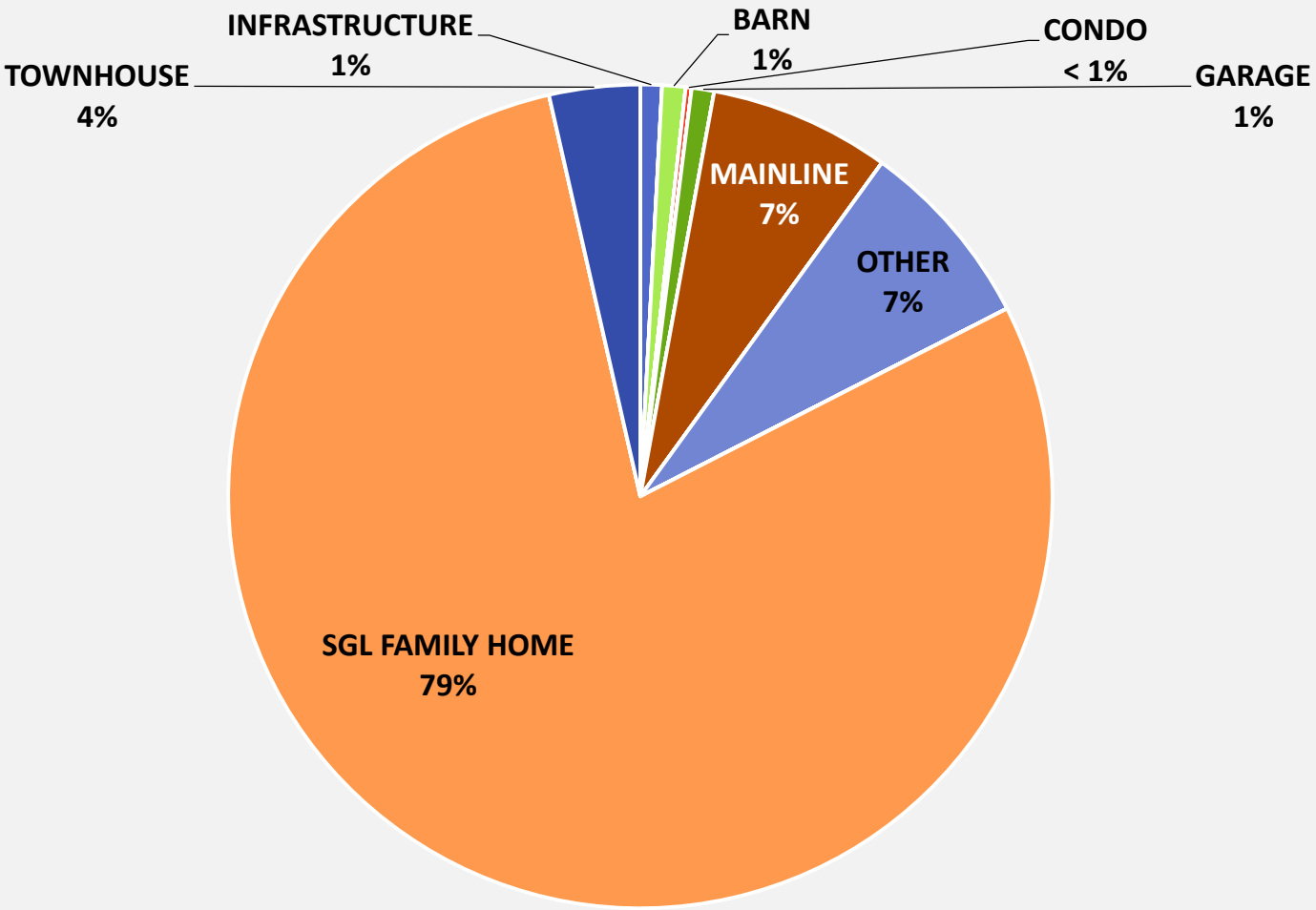
Data Description

- Data for each construction project stored in a Work Management System (WMS) database
- Time frame of data spans 10 years (2005 - 2015)
- Approximately 300,000 data points in a comma-separated values (CSV) file
- Data fields includes:
  - Work Number
  - Job Classification (Dwelling Types, Mainline, Infrastructure, etc.)
  - Construction Classification (Length of Conductor Cable, Trenching, Transformers, etc.)
  - Cost of construction (Material, Labor, and Overhead)

Data Scrubbing

- Filtered all residential related data
- Combined rows with the same work number which is assumed to belong to a single project
- Used the number of meters for each project to keep track of the number of homes
- Added Infrastructure classification which includes Cable TV, HOA Light, Traffic Signal, etc.
- This reduced the data file from 300,000 lines to approximately 25,000 lines

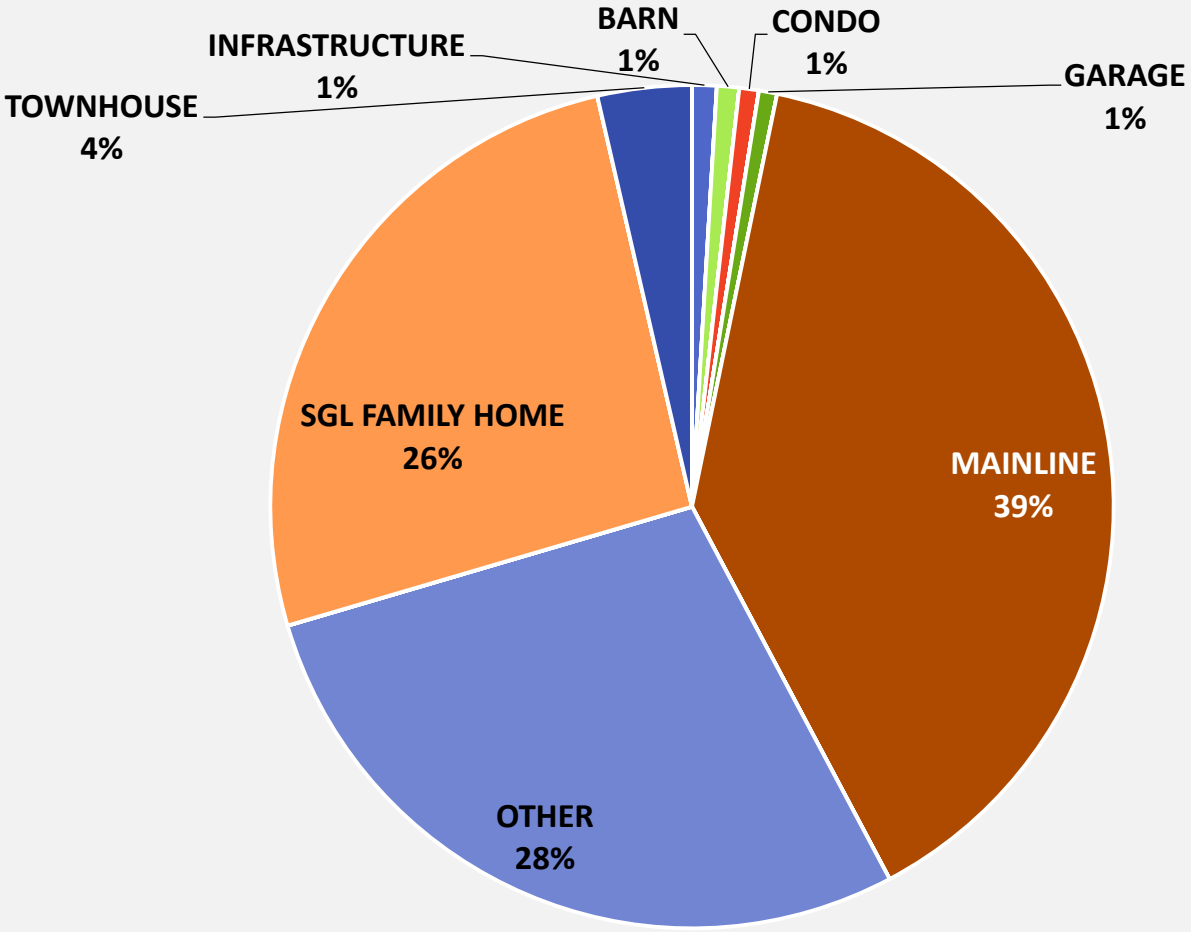
Percentage of Job Counts by Job Classification



Cost Data Normalization

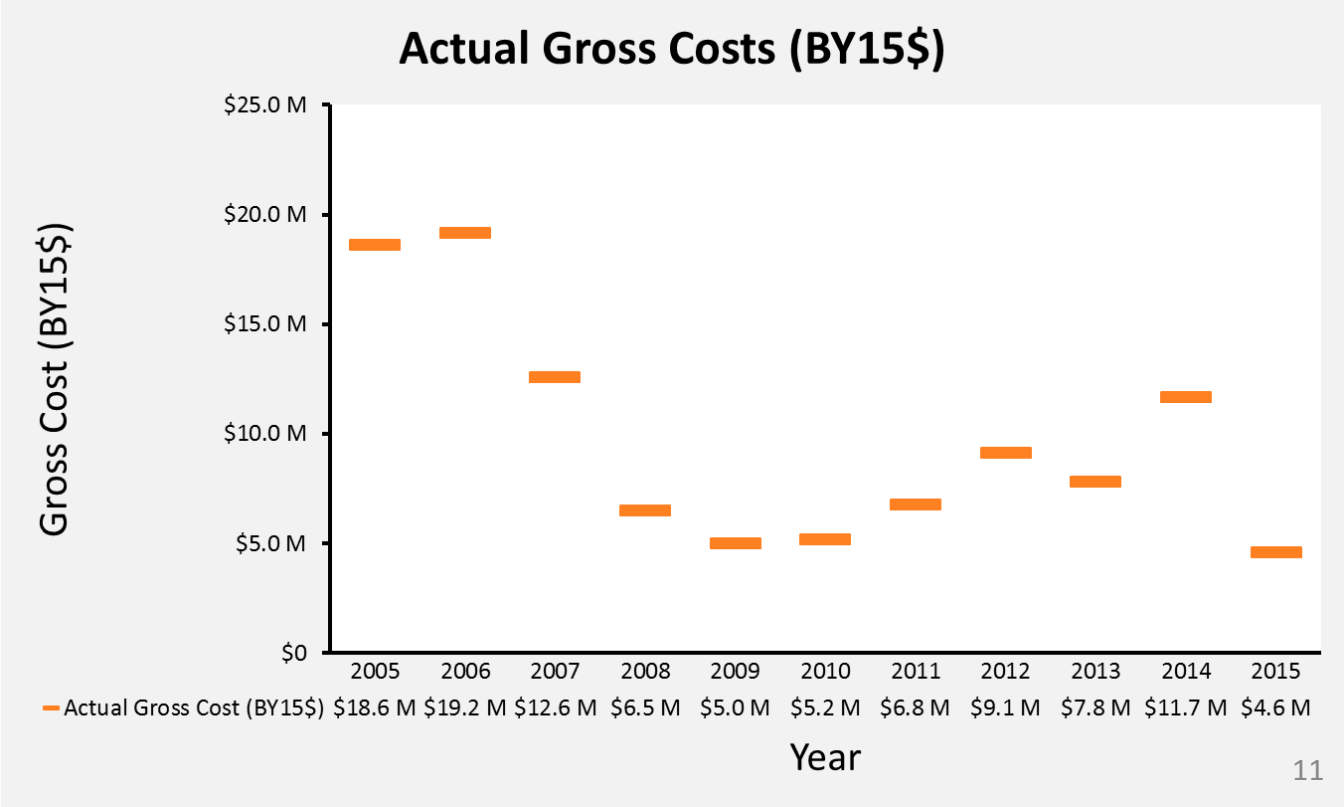
- All costs assumed to have been recorded in the base year of the project completion year. This assumption was confirmed by the client.
- Utilized Handy Whitman Index of Construction Cost to account for inflation
  - Isolated data for Electricity Utility Construction for Distribution Plants for the South Atlantic Region
  - Analyzed cost changes in goods related to electric utility construction to calculate annual inflation normalization factors

Percentage of Gross Costs by Job Classification



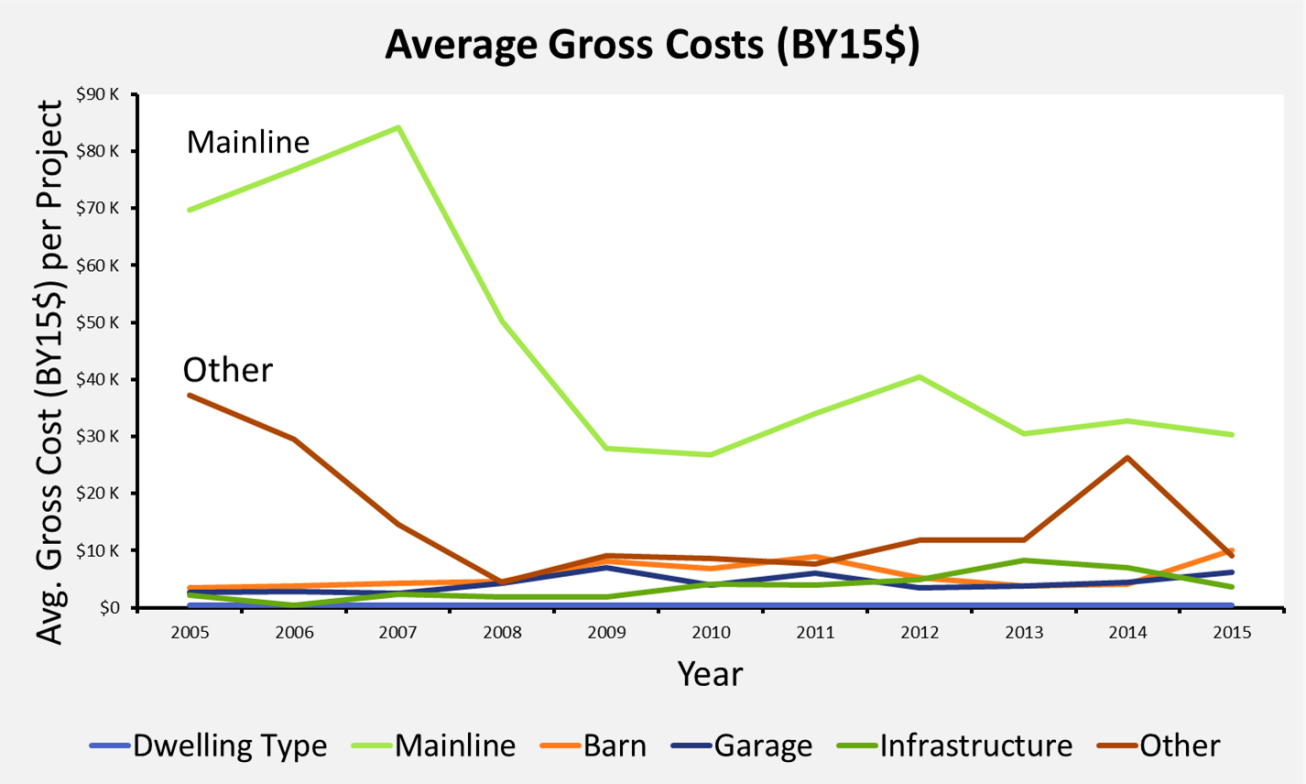
Data Selection

- Gross costs from 2005 - 2007 exhibits different behavior than the costs from 2008 - 2014
- Relatively high number of expensive projects in the years of 2005 - 2007
- The phenomenon is assumed to be caused by the performance of housing market and economy



Data Selection

- The average costs for mainline and other projects are relative high in 2005 - 2007
- Other types of construction do not have huge variances
- The cost trend from 2005 - 2007 is not suitable to predict cost in short term
- Data points from 2005 - 2007 were removed for further analysis



Modeling

Regression

- Utilize technical parameters to predict total gross cost
- Technical parameters associated with home type jobs (Single Family Homes, Townhomes, and Condos)
- Assume costs for Mainline, Infrastructure, etc. as “cost of doing business”

Simulation

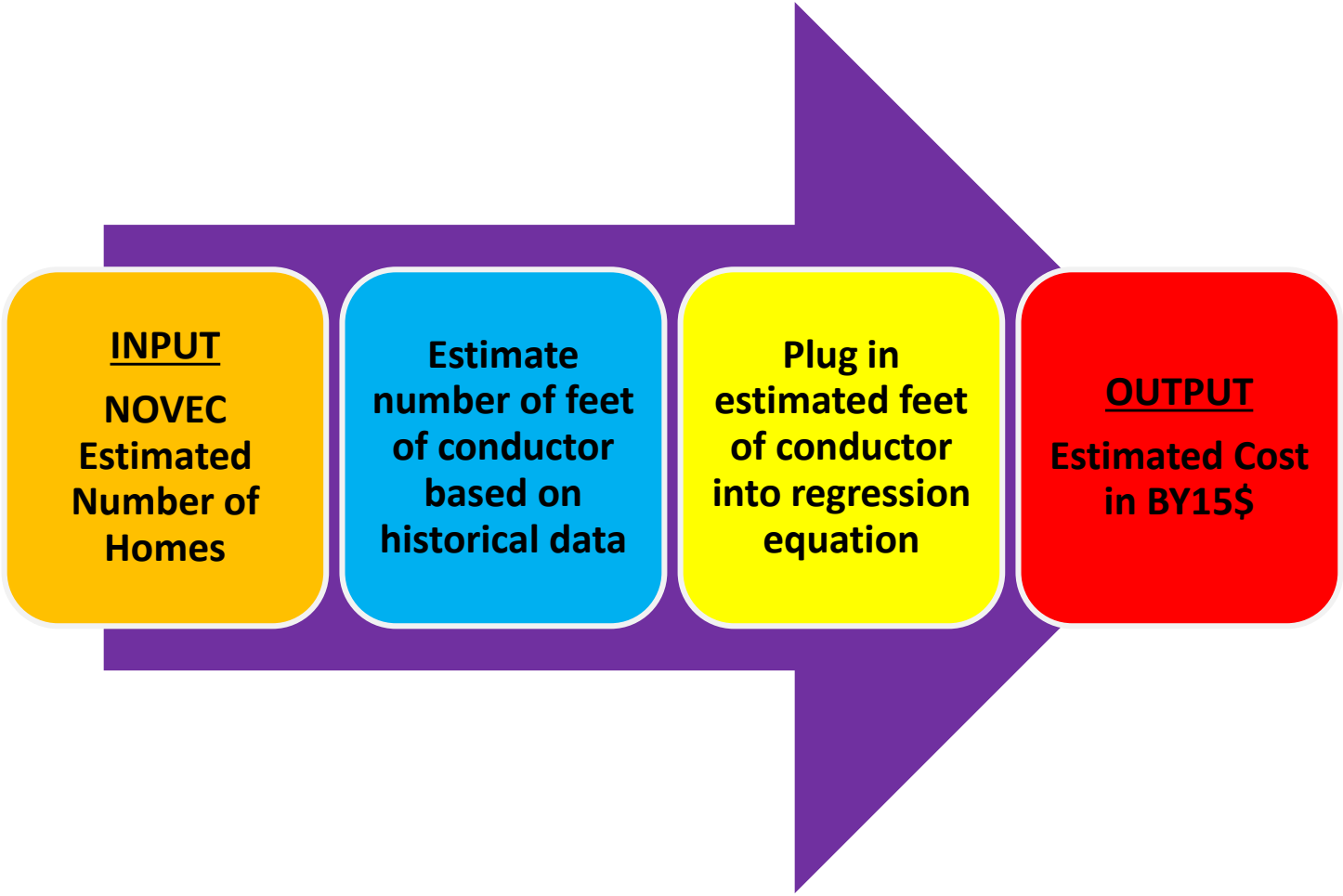
- Apply bootstrapping and Monte Carlo simulation method to predict the total cost
- Utilize all project types (unlike regression model)
- Use number of houses as the input and predict number of each product type by
  - Regression (for Mainline)
  - ARIMA time series method (for Barn, Garage, Infrastructure, Other)
- Run simulation to predict cost of each category from the historical data set

Regression Model

Regression Model Approach

- Leveraged R Statistical language to calibrate model
- Investigated and analyzed several types and combinations of regressions and predictor variables
- Selected length of conductor cable parameter due to it being a major cost driver, solid linear relationship to baseline job gross cost, and high correlation
- Cross validated model

Regression Model Algorithm

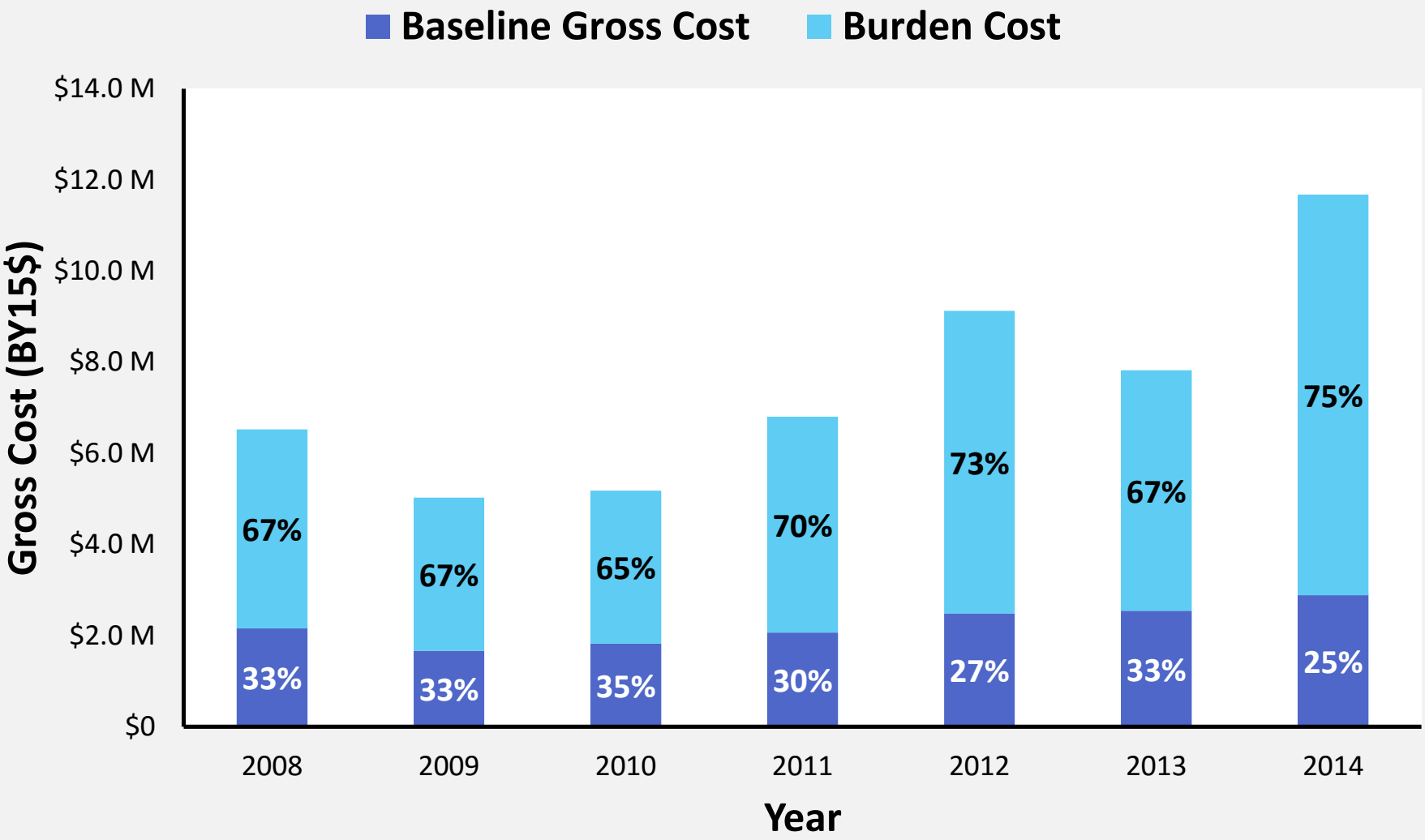


Burden Costs

- Baseline Gross Costs
  - Single-Family Home
  - Townhome
  - Condo

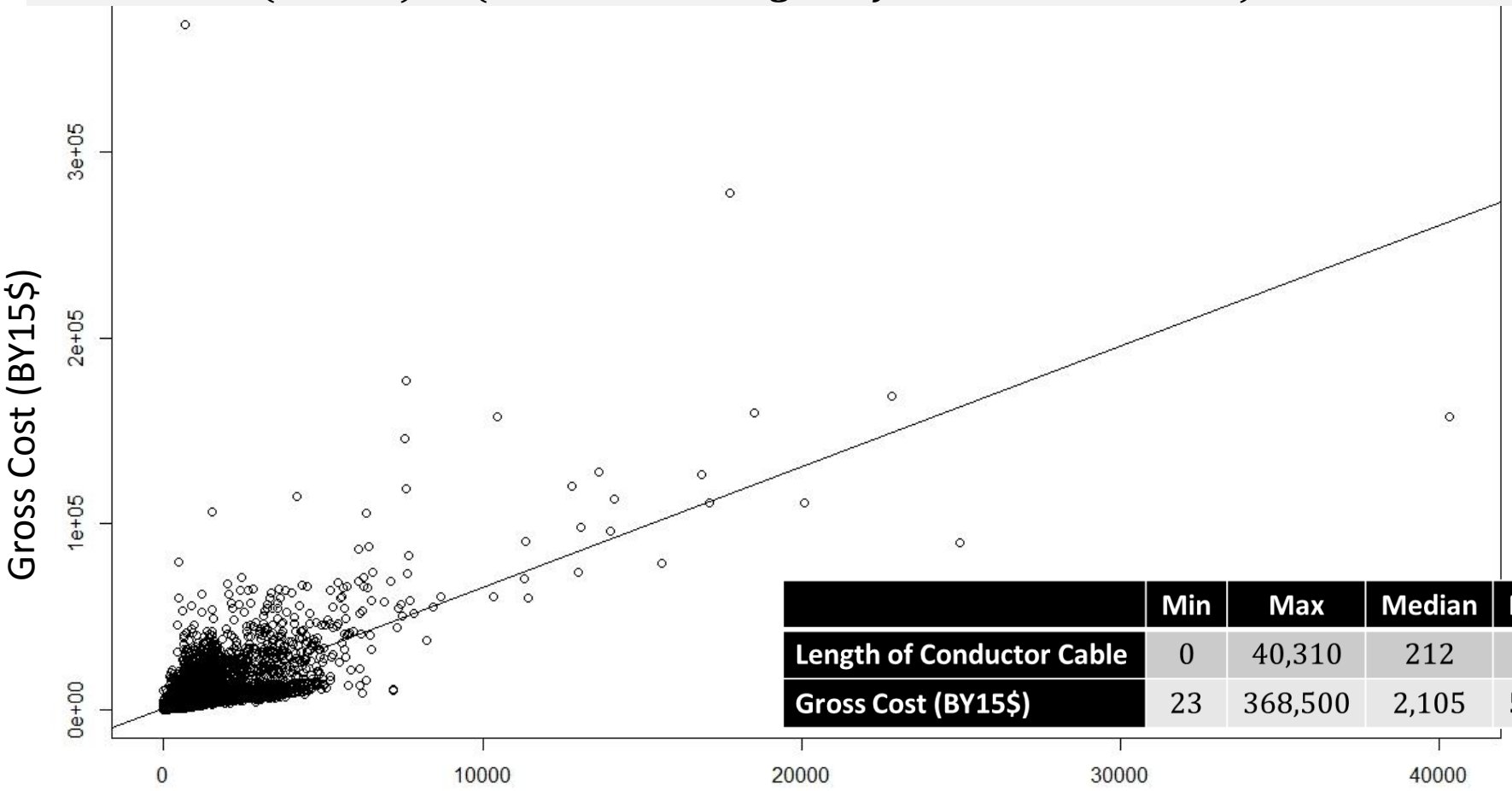
- Burden Costs
  - Mainline
  - “Other”
  - Barn
  - Garage
  - Infrastructure

Burden Costs Cont.



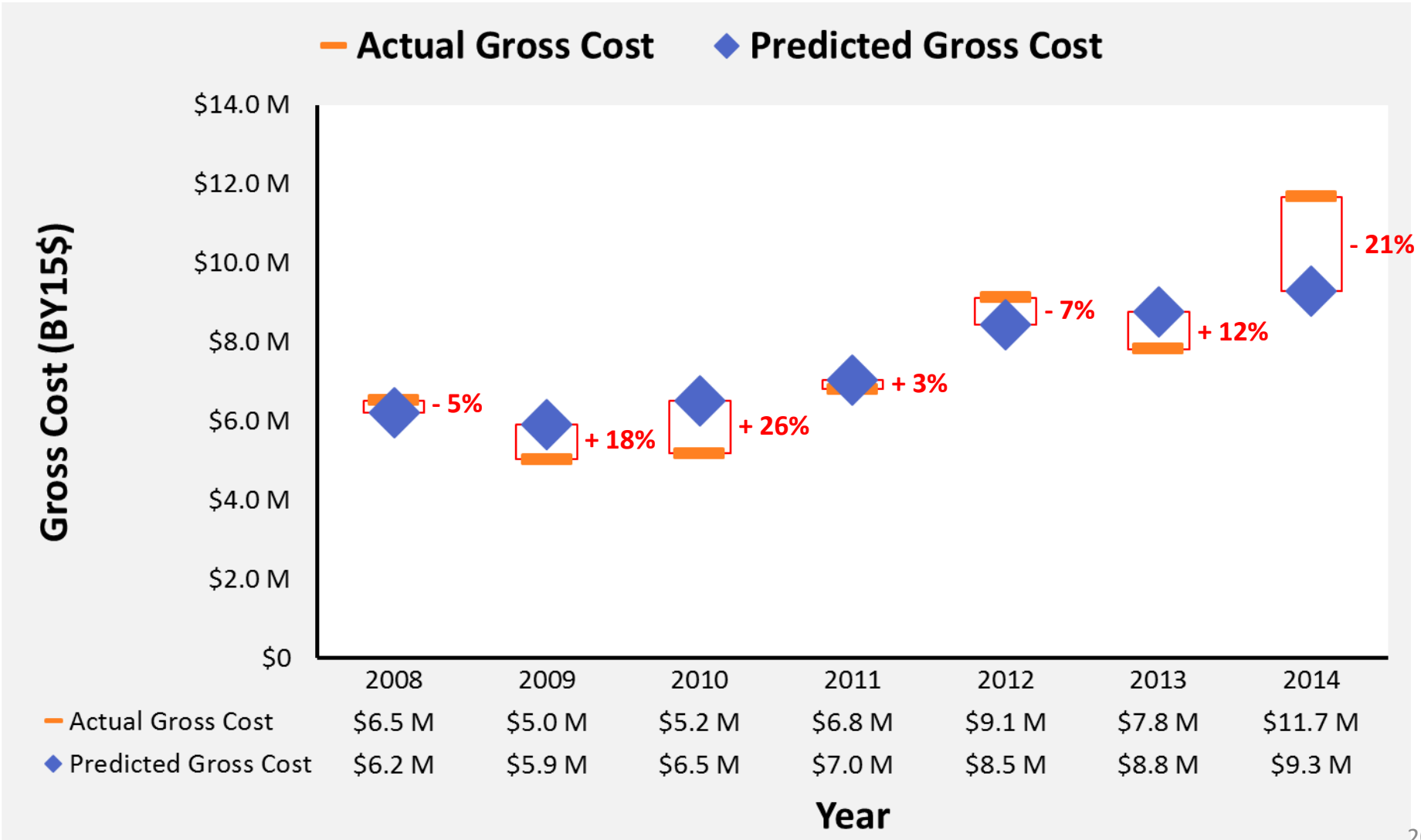
Linear Regression
Gross Cost vs Conductor Cable

$$\textbf{Gross Cost (BY15\$)} = (6.48475 \times \textbf{Length of Conductor Cable}) + 918.12091$$

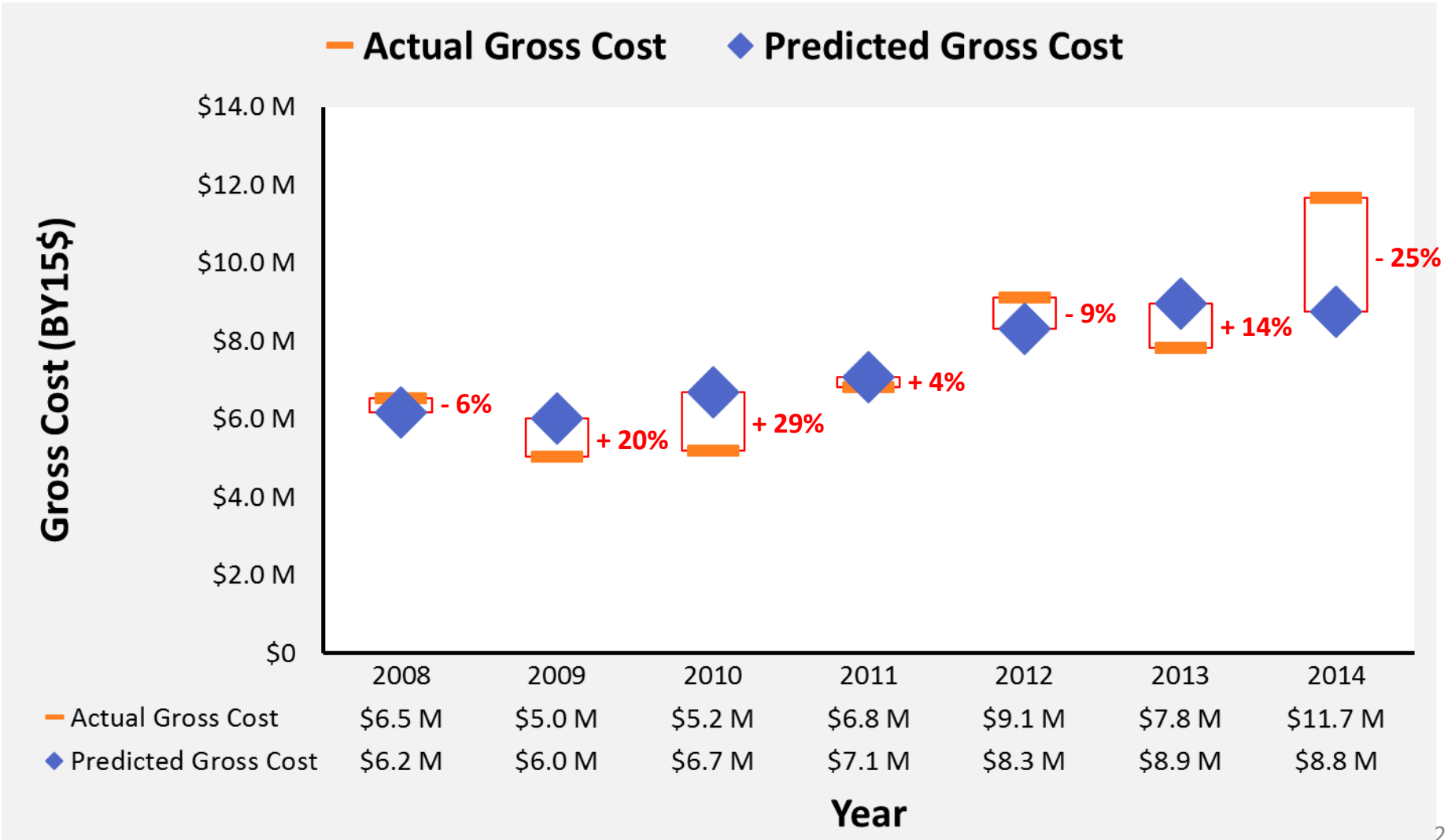


	Min	Max	Median	Mean
Length of Conductor Cable	0	40,310	212	586
Gross Cost (BY15$)	23	368,500	2,105	5,106

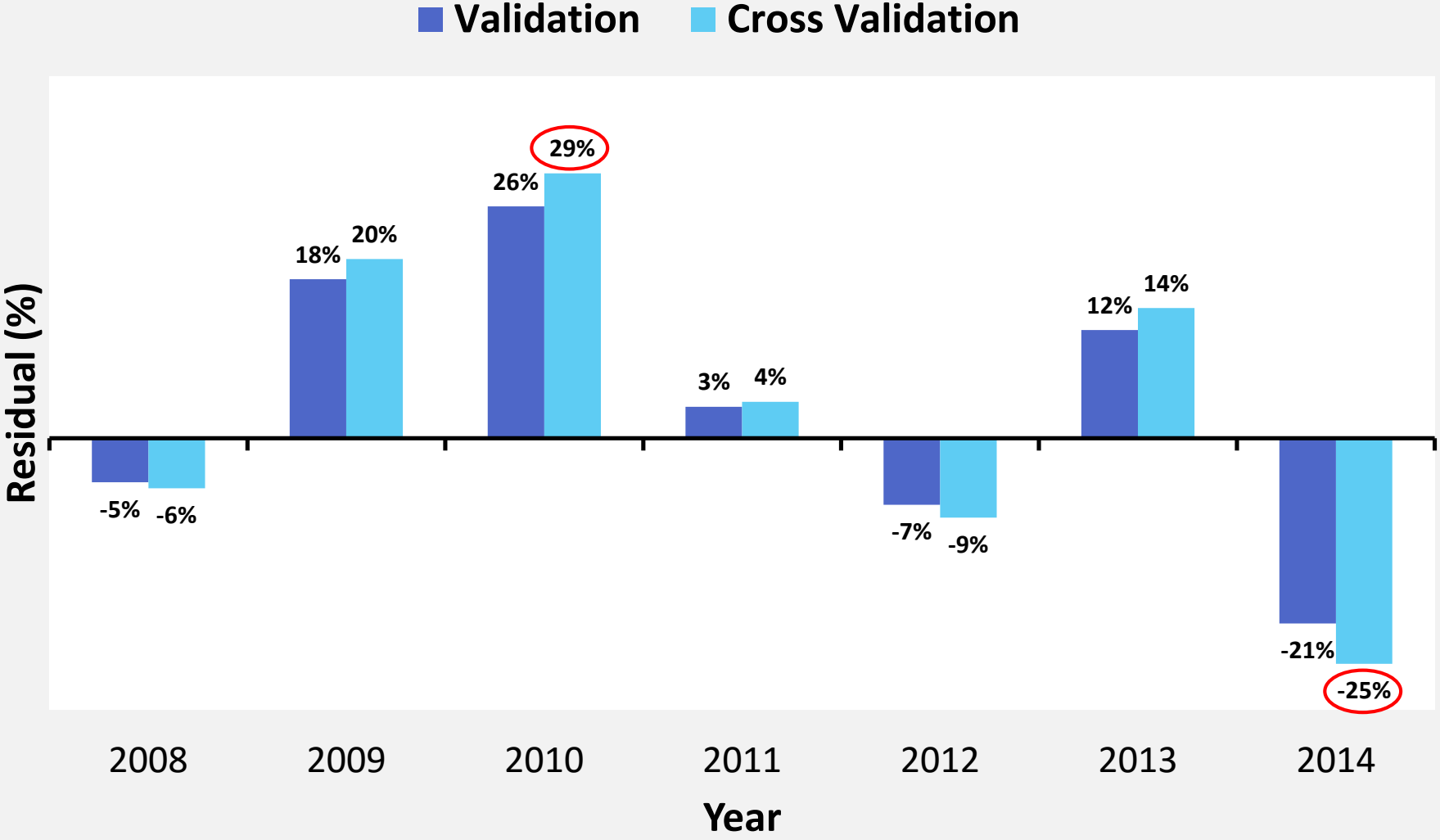
Regression Model Initial Validation



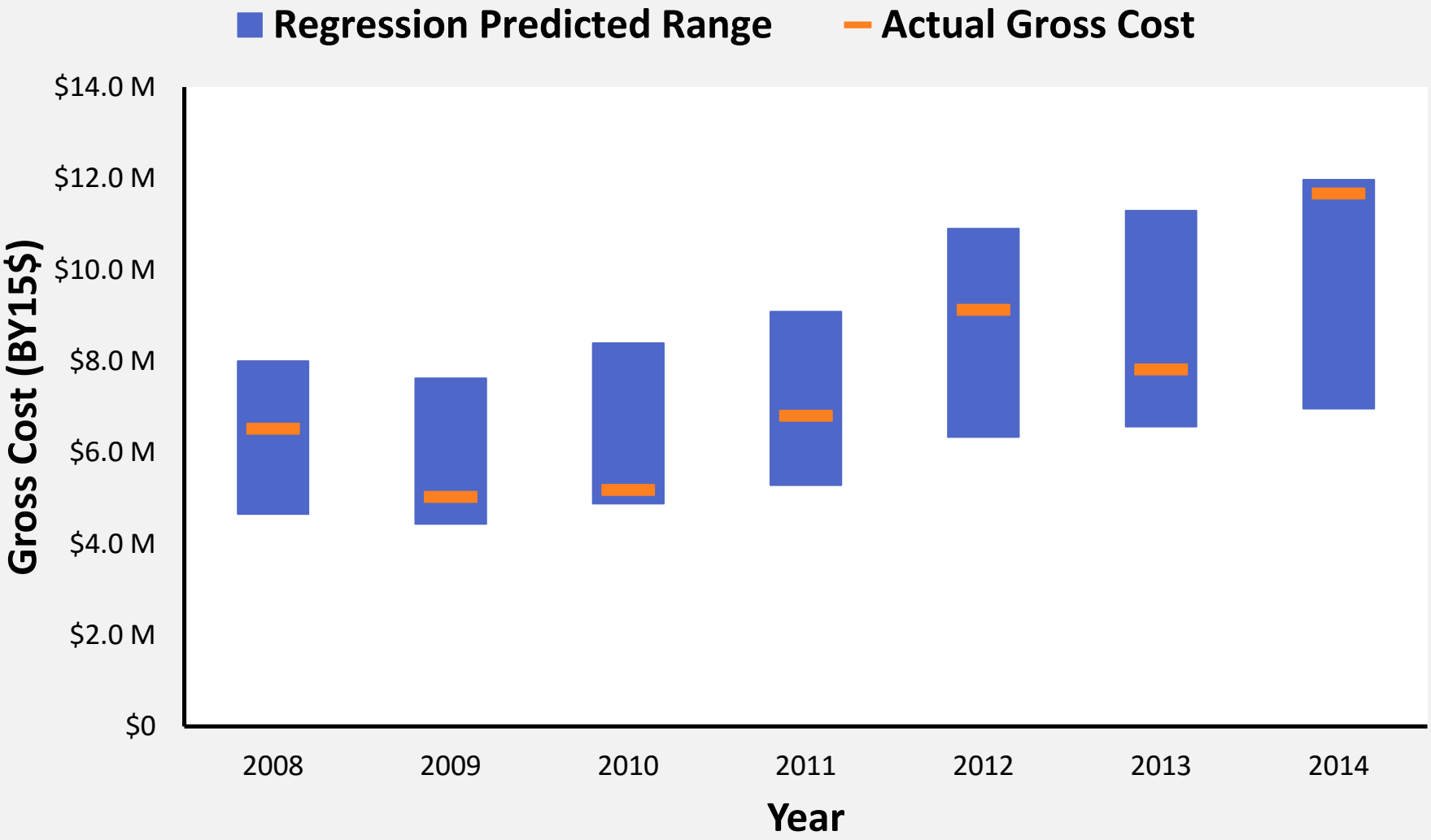
Regression Model Cross Validation



Residual Comparison



Regression Model Predicted Range

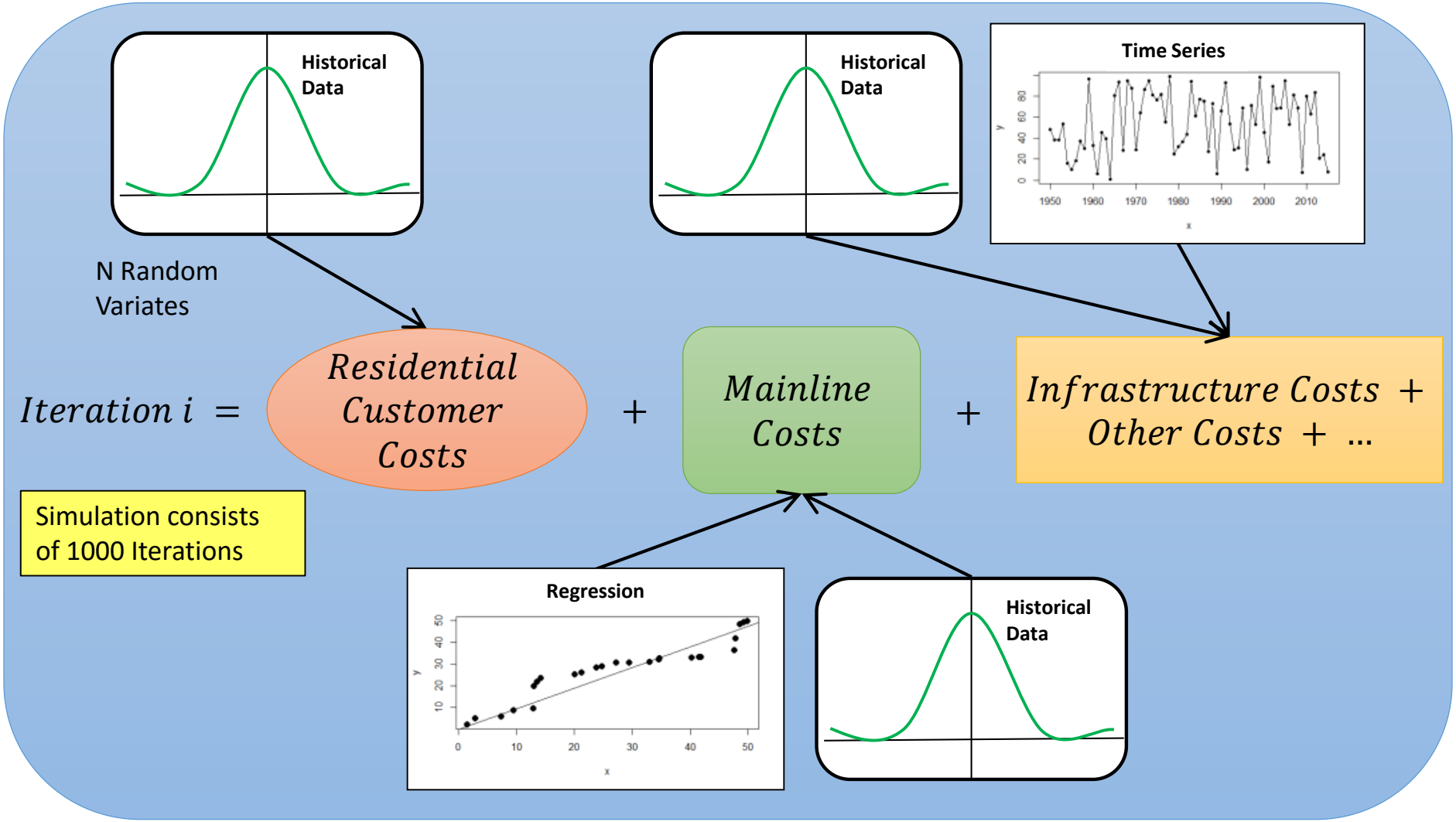


Stochastic Simulation Model

Simulation Model Algorithm

- Input: Total Number of Houses
- For one iteration
  - Direct Cost: Cost for Houses
    - Distribute the cost per house for all the types of houses
    - Convert total number of houses into different types
    - Bootstrap costs based on the predicted number of houses
    - Sum the total cost for houses
  - Indirect Cost: Cost for Other Types
    - Find relationship between number of projects in each category and number of houses
    - Predict number of projects in each category
    - Bootstrap costs based on the number of projects in each category
    - Sum the total cost for each category
  - Sum All the Costs
- Run Multiple Iterations

Simulation Model Algorithm



Number of Non-Dwelling Jobs

- High correlation between number of homes and number of mainline jobs
  - Linear Regression (R-square = 0.907)
  - $Number\ of\ Mainline\ Jobs = (0.0330 \times Number\ of\ homes) - 62.15$
- Low correlation between number of homes and number of non-dwelling jobs except mainline jobs
  - Use Time Series Estimation (ARIMA Method)

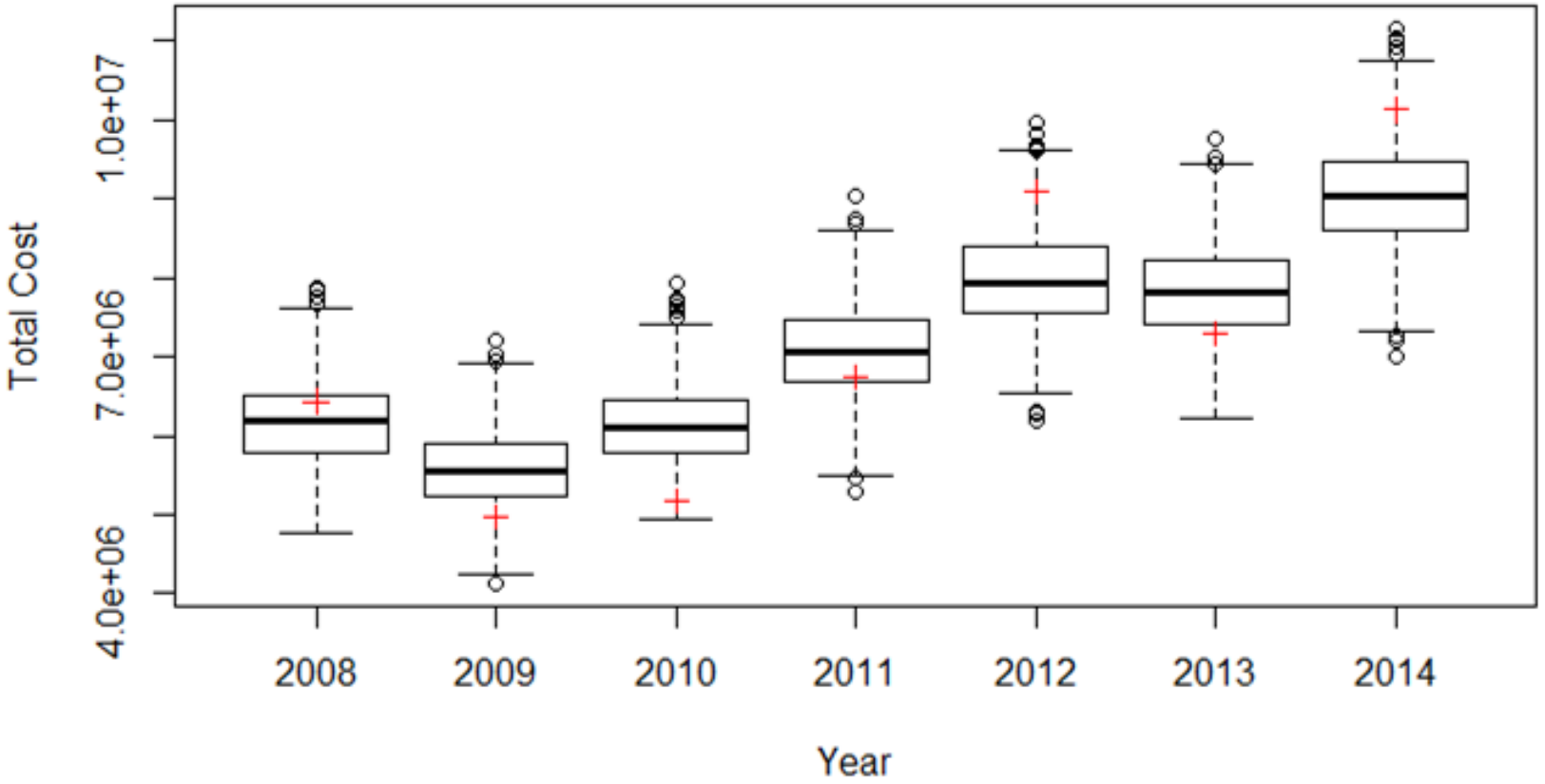
Job Classification	Total Number of Homes
Barn	0.310
Garage	-0.483
Infrastructure	-0.0873
Other	-0.314
Mainline	0.952

Simulation Cross Validation

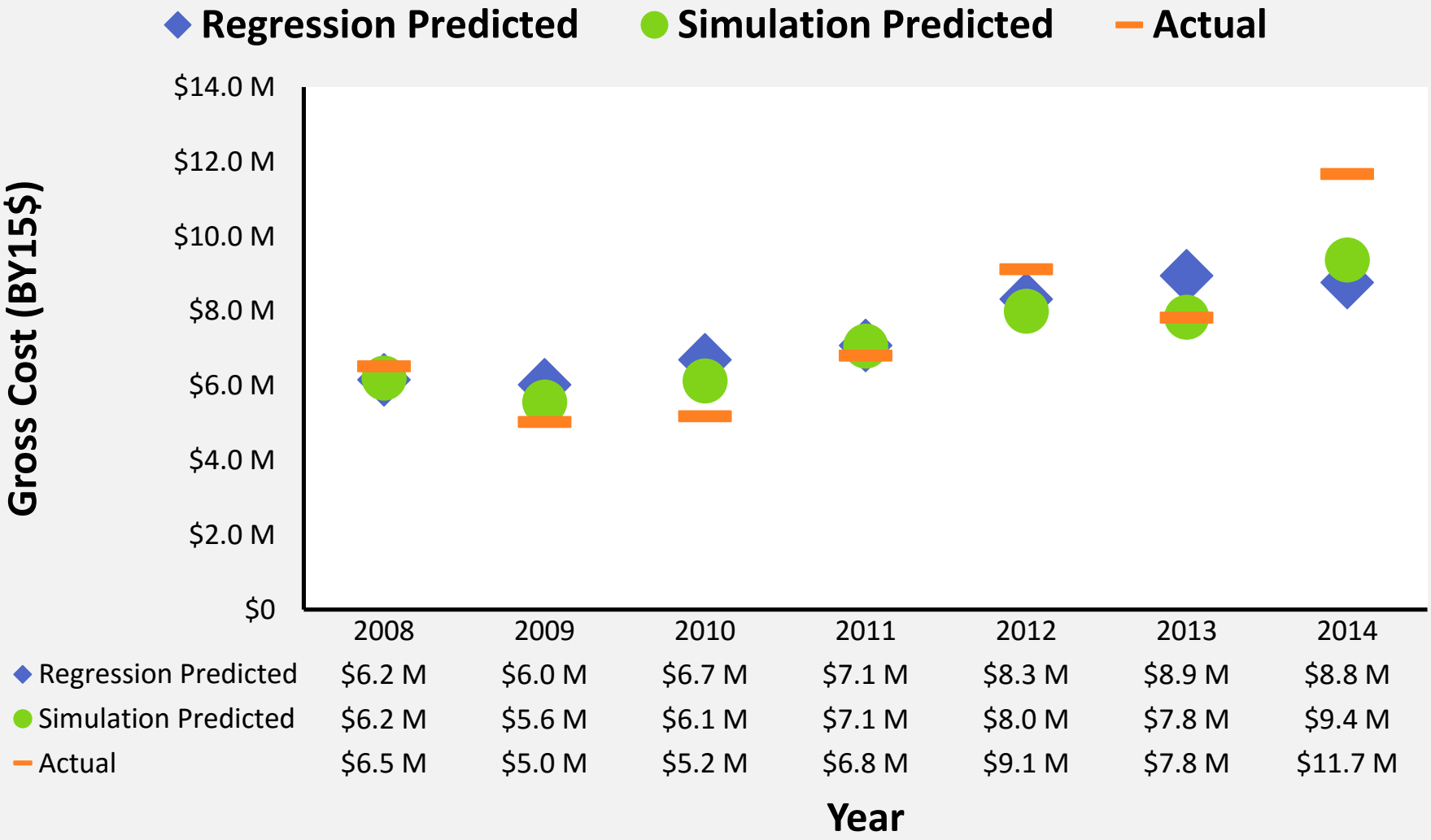
Year	Actual Cost (BY15$)	Average Predicted Cost (BY15$)	Percent Error
2008	6.44 M	6.21 M	- 3.56%
2009	4.96 M	5.56 M	+ 12.13%
2010	5.18 M	6.13 M	+ 18.30%
2011	6.74 M	7.06 M	+ 4.85%
2012	9.10 M	8.00 M	- 12.10%
2013	7.29 M	7.84 M	+ 7.54%
2014	10.1 M	9.37 M	- 7.43%

Simulation Cross Validation

- Actual Cost vs Simulation Outcomes



Comparison



	2008	2009	2010	2011	2012	2013	2014
Regression Predicted	$6.2 M	$6.0 M	$6.7 M	$7.1 M	$8.3 M	$8.9 M	$8.8 M
Simulation Predicted	$6.2 M	$5.6 M	$6.1 M	$7.1 M	$8.0 M	$7.8 M	$9.4 M
Actual	$6.5 M	$5.0 M	$5.2 M	$6.8 M	$9.1 M	$7.8 M	$11.7 M

Comparison Cont.

- Consistent trend between Regression and Simulation
- Heuristic approach:
  - Overestimating:
    - Regression Predicted Gross Cost **>** Simulation Predicted Gross Cost
    - Actual cost should be lower than Simulation
  - Underestimating:
    - Regression Predicted Gross Cost **<** Simulation Predicted Gross Cost
    - Actual cost should be higher than Simulation

	-	+	+	+	-	+	-
Regression	- 6%	+ 20%	+ 29%	+ 4%	- 9%	+ 14	- 25%
Simulation	- 4%	+ 12%	+ 18%	+ 5%	- 12%	+ 8%	- 7%
Year	2008	2009	2010	2011	2012	2013	2014

Recommendations

- Investigate if it is feasible to allocate Mainline and Infrastructure costs to the particular jobs or development that the element is supplying/supporting
  - Allows for accurate allocation of “cost of business” to home counts and should aid in using home counts as a predictor of cost
  - Consider adding number of Mainline jobs as a predictor to future models
- Incorporate more consistent data recording
  - Accurate labeling of Units of Measure, Quantities, and IDs necessary to eliminate fundamental differences between what should be similar expense items
  - Eliminate redundancy of recorded elements (e.g. meters, labor,...)
- For the long term, implementation of a GIS system to know where all of their assets reside and to provide linkage between components
  - Allows for advanced spatial analysis
  - Benefits maintenance department
  - Record latitude and longitude (short term solution)

Way Forward

- Identify and document lead and lagging indicators for the linking of development construction and Mainline and Infrastructure jobs
  - Effects occur in following years
  - Eliminates the need for allocating costs to baseline job costs by burdening
- Investigate “Other” Job classifications
  - “Other” jobs are a significant source of cost
  - May contain both commercial and residential costs
  - Understanding what these are may lead to classification change or better alignment with existing Job Work Number
- Consider cost variances among jobs with similar attributes
- Analyze effects from terrain and working environments
  - Plain/hills
  - Overhead/underground

Acknowledgements

NOVEC

- Bryan Barfield
- Shayan Rashid

GMU

- Dr. Karla Hoffman
- Dr. Kuo-Chu Chang
- Dr. Jie Xu

Questions?