Designing Smart-grid Telecommunications Systems via Interval Flow Network Optimization

Anthony J. Klinkert
Southern Methodist University, klinkert@smu.edu

Follow this and additional works at: https://scholar.smu.edu/engineering_management_etds

Part of the Systems and Communications Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Operations Research and Engineering Management at SMU Scholar. It has been accepted for inclusion in Operations Research and Engineering Management Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
DESIGNING SMART-GRID TELECOMMUNICATIONS
SYSTEMS VIA INTERVAL FLOW NETWORK
OPTIMIZATION

Approved by:

______________________________
Richard S. Barr

______________________________
Eli V. Olinick

______________________________
Michael Hahsler

______________________________
M. Scott Kingsley

______________________________
Robert H. Jones
DESIGNING SMART-GRID TELECOMMUNICATIONS SYSTEMS VIA INTERVAL FLOW NETWORK OPTIMIZATION

A Praxis Presented to the Graduate Faculty of the Bobby B. Lyle School of Engineering Southern Methodist University in Partial Fulfillment of the Requirements for the degree of Doctor of Engineering with a Major in Engineering Management by

Anthony Jacob Klinkert

(B.S.E.E., University of Texas at Austin, 1979) (M.S.E.E., Southern Methodist University, 1984) (M.S. Telecommunications, Southern Methodist University, 1999)

May 19, 2018
ACKNOWLEDGMENTS

I would like to thank the EMIS faculty, staff and my praxis committee at SMU for their support throughout the D.Eng. process. I would especially like to acknowledge and thank Dr. Richard S. Barr for his guidance, insight and suggestions.
Abstract

One of the first major, widespread deployments of an Internet of Things (IoT) system is an electric utility’s smart-grid communications network. Communication networks support utility smart-grids for an electrical power provider by linking industrial IoT devices on the grid, such as residential and commercial customers’ electric meters to the utility’s data center. A new solution methodology and software tool is applied to the optimization of an electric utility smart-grid communications network. This research provides utilities with a technique to reduce cost of new and current communications networks for smart-grid programs.

This research introduces the use of a new advanced optimization heuristic and solver effective for this problem. Interval-flow networks and associated solution methods offer a new approach to modeling and optimization that can readily capture the hierarchical structure of these minimum-cost network-topology optimization problems, and can quickly solve large-scale network instances. Computational experiments provide insight into the effect that key situational and practitioner decisions have on
resulting designs. This work promises a methodology and tool useful to practitioners responsible for smart-grid communications network realization.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS .................................................................................. iii

List of Figures .......................................................................................... xii

List of Tables ............................................................................................. xv

LIST OF ABBREVIATIONS ......................................................................... xix

CHAPTER

1. INTRODUCTION AND OVERVIEW ......................................................... 1
   1.1. Introduction ...................................................................................... 1
   1.2. Power-Engineering Context ............................................................ 3
   1.3. The Problem .................................................................................... 9
   1.4. Literature Review ............................................................................. 10
       1.4.1. Engineering Management Literature .................................... 10
       1.4.2. Electrical Engineering Literature ........................................ 11
       1.4.3. Operations Research Literature ............................................ 13
       1.4.4. Power-Systems Engineering and SGCN Design Literature ... 14
       1.4.5. SGCN Design Literature ...................................................... 16
       1.4.6. Experimental Design Literature .......................................... 17
   1.5. The Solution Approach ................................................................ 18
   1.6. Expected Contributions ................................................................ 21

2. PROBLEM DEFINITION AND MODELLING APPROACH ............. 23
   2.1. Problem Definition ......................................................................... 23
   2.2. Mathematical Notation, Conventions and Definitions ............... 24
2.3. Formulation of a SGCN Design Optimization Model................. 25
2.3.1. Uncapacitated Minimum Cost Network Flow Model ........ 35
2.3.2. Capacitated Minimum Cost Network Flow Model ......... 38
2.3.3. Fixed Charge Capacitated Minimum Cost Network Flow Model ................................. 39
2.3.4. Conditional Lower Bounds and Interval Flow Networks ... 41
2.3.5. Fixed-charge Interval-Flow Network Model ................. 42
2.3.6. Formulating the Complete SGCN Model .................. 42
2.4. The SGCN Model ........................................ 49

3. EXPERIMENTATION ......................................... 55
3.1. Experimental Test Problems .................................. 55
3.2. Response Variables and Performance Evaluation Criteria .... 56
3.2.1. Total Network Cost ..................................... 57
3.2.2. Equipment Count ........................................ 58
3.2.3. Average Link Distance .................................. 58
3.3. Factors and Levels to be Explored ......................... 59
3.4. The Hypothesis to Be Tested ................................. 61
3.5. The Design .................................................. 63
3.6. The Experimental Scenario .................................. 67
3.7. The Analysis ................................................ 72
3.8. The Experimental Data ...................................... 72
3.9. Interpretation of Analysis Computations ..................... 75
3.10. Analysis of Each Response in the Experiment .............. 83
3.10.1. Total Network Cost ...................................... 83
3.10.2.1.2 The Influence of Area on Equipment Count .......................... 118
3.10.2.1.3 The Influence of Lower-Bound on Equipment Count .................. 121
3.10.2.1.4 The Influence of Equipment-Price on Equipment Count ................. 124

3.10.2.2. Pair-wise Analysis .................................................... 127
3.10.2.2.1 The Influence of Meters*Area on Equipment Count ..................... 127
3.10.2.2.2 The Influence of Meters*Lower-Bound on Equipment Count .............. 128
3.10.2.2.3 The Influence of Meters*Equipment-Price on Equipment Count ........... 130
3.10.2.2.4 The Influence of Area*Lower-Bound on Equipment Count ................. 132
3.10.2.2.5 The Influence of Area*Equipment-Price on Equipment Count ............. 133
3.10.2.2.6 The Influence of Lower-Bound*Equipment-Price on Equipment Count .... 135

3.10.2.3. Equipment Count Regression Analysis ................................. 137
3.10.2.4. Summary for Equipment Count .......................................... 141

3.10.3. Average Link Distance ................................................... 142
3.10.3.1. Single-Factor Analysis .................................................. 142
3.10.3.1.1 The Influence of Meters on Average Link Distance ....................... 143
3.10.3.1.2 The Influence of Area on Average Link Distance ........................ 146
3.10.3.1.3 The Influence of Lower-Bound on Average Link Distance ................. 148
3.10.3.1.4 The Influence of Equipment-Price on Average Link Distance ............... 151

3.10.3.2 Pair-wise Analysis .................................................. 154

3.10.3.2.1 The Influence of Meters*Area on Average Link Distance ................ 154

3.10.3.2.2 The Influence of Meters*Lower-Bound on Average Link Distance ....... 157

3.10.3.2.3 The Influence of Meters*Equipment-Price on Average Link Distance .... 158

3.10.3.2.4 The Influence of Area*Lower-Bound on Average Link Distance .......... 160

3.10.3.2.5 The Influence of Area*Equipment-Price on Average Link Distance ...... 162

3.10.3.2.6 The Influence of Lower-Bound*Equipment-Price on Average Link Distance .... 163

3.10.3.3 Average Link Distance Regression Analysis ............... 165

3.10.3.4 Summary for Average Link Distance ......................... 169

3.10.4 Summary Statistical Analysis ....................................... 171

3.11. Findings ................................................................. 175

3.11.1 Total Network Cost .................................................. 176

3.11.2 Equipment Count ..................................................... 177

3.11.3 Average Link Distance ............................................... 178

3.11.4 Summary Findings ..................................................... 179

4. CONCLUSIONS AND FUTURE RESEARCH ................................ 180

4.1 Summary of Findings and Conclusion .................................. 180

4.2 Contributions of the Research ......................................... 181

4.3 Future Research .......................................................... 182
## List of Figures

1.1 Smart-grid Power and Communications Flows (Source: [53]) ............... 3
1.2 Geographic Expanse of the Smart-Grid (Source: [45]) .................. 6
1.3 Applications and Infrastructure of SGCNs (Source: [45]) .............. 7
2.1 Model of a Communications Arc ........................................ 29
2.2 Transportation Network Model .......................................... 32
2.3 Example Transshipment Network - Transshipment Nodes Shown in Blue .......................................................... 37
2.4 Simple Example Smart-meter Network ..................................... 43
2.5 Split Node ................................................................. 45
2.6 Fully Displayed Arcs .......................................................... 46
2.7 Concisely Displayed Arcs .................................................. 46
2.8 SGCN Model Graph .......................................................... 49
2.9 SGCN Model Graph with Sets ............................................ 51
3.3 Household Density, Dallas Texas USA ..................................... 69
3.4 Suburban Neighborhood Example ......................................... 70
3.5 LS Means Plot Gender vs. Pain ........................................... 76
3.6 LS Means Plot of Meters vs. Total Network Cost ....................... 84
3.7 LS Means Plot Area vs. Total Network Cost .............................. 88
3.8 LS Means Plot Lower-Bound vs. Total Network Cost ................. 90
3.9 LS Means Plot Equipment-Price vs. Total Network Cost ............. 93
3.10 LS Means Plot Meters*Area Interaction vs. Total Network Cost . . . 96
3.11 LS Means Plot Meters*Lower-Bound, vs. Total Network Cost ........ 99
3.12 LS Means Plot Meters*Equipment-Price Interaction vs. Total Network Cost .......................................................... 101
3.13 LS Means Plot Area*Lower-Bound vs. Total Network Cost ......................... 103
3.14 LS Means Plot Area*Equipment-Price Interaction vs. Total Network Cost 105
3.15 LS Means Plot Lower-Bound*Equipment-Price vs. Total Network Cost . 107
3.16 Actual by Predicted Plot .......................................................... 111
3.17 Prediction Model for Total Network Cost ........................................ 112
3.18 Means Plot Meters vs. Equipment Count ..................................... 116
3.19 LS Means Plot Area vs. Equipment Count .................................... 119
3.20 LS Means Plot Lower-Bound vs. Equipment Count .......................... 122
3.21 LS Means Plot Equipment-Price vs. Equipment Count ..................... 125
3.22 LS Means Plot Meters*Area Interaction vs. Equipment Count .......... 128
3.23 LS Means Plot Equipment-Price vs. Equipment Count ................... 129
3.24 LS Means Plot Meters*Equipment-Price Interaction vs. Equipment Count .......................................................... 131
3.25 LS Means Plot Area*Lower-Bound vs. Equipment Count .................. 133
3.26 LS Means Plot Area*Equipment-Price Interaction vs. Equipment Count 134
3.27 LS Means Plot Lower-Bound*Equipment-Price vs. Equipment Count . 136
3.28 Actual by Predicted Plot Equipment Count .................................... 139
3.29 Prediction Model for Equipment Count ........................................ 140
3.30 LS Means Plot Average Link Distance vs. Average Link Distance ...... 144
3.31 LS Means Plot Area vs. Average Link Distance .............................. 147
3.32 LS Means Plot Lower-Bound vs. Average Link Distance ................... 149
3.33 LS Means Plot Lower-Bound vs. Average Link Distance ................... 152
3.34 LS Means Plot Meters*Area Interaction vs. Average Link Distance ... 155
3.35 LS Means Plot Meters*Lower-Bound vs. Average Link Distance ........ 158
3.36 LS Means Plot Meters*Equipment-Price Interaction vs. Average Link Distance .................................................. 159
3.37 LS Means Plot Area*Lower-Bound vs. Average Link Distance ........ 161
3.38 LS Means Plot Area*Equipment-Price Interaction vs. Average Link Distance ........................................................... 162
3.39 LS Means Plot Lower-Bound*Equipment-Price vs. Average Link Distance 164
3.40 Actual by Predicted Plot Average Link Distance ......................... 168
3.41 Prediction Model for Average Link Distance ......................... 169
### List of Tables

1.1 Smart-grid Functionalities and Communications Needs (Source: [23]) ... 8

3.3 Hypotheses to be Tested .......................................................... 62

3.4 Calculation of Areas in Square Miles ........................................ 68

3.5 Densities ................................................................................. 68

3.6 Solver Parameter Inputs ............................................................. 71

3.7 Experimental Data ................................................................. 73

3.8 Experiment Variables - Main Effect and Interaction Terms .......... 74

3.9 Experiment Variables - Responses ............................................. 75

3.10 Example ANOVA table .............................................................. 76

3.11 Student’s $t$-test for Meters Comparing 10,000 with 5,000 ......... 79

3.12 Letters Report Meters ............................................................... 80

3.13 Sorted Parameter Estimates ..................................................... 81

3.14 ANOVA Table ........................................................................... 82

3.15 Summary of Fit Table .............................................................. 82

3.16 ANOVA for Meters ................................................................. 85

3.17 Student’s $t$-test for Meters Comparing 10,000 with 5,000 ......... 86

3.18 Letters Report Meters ............................................................... 86

3.19 ANOVA for Area ................................................................. 88

3.20 Student’s $t$-test for Area Comparing 500 with 100 Areas .......... 89

3.21 Letters Report Area ............................................................... 89

3.22 ANOVA for Lower-Bound .................................................... 91

3.23 Student’s $t$-test for Lower-Bound Comparing 200 with 20 ....... 91

xv
3.24 Letters Report Lower-Bound ........................................ 92
3.25 ANOVA for Equipment-Price ........................................ 93
3.26 Student’s t-test for Equipment-Price Comparing 200,000 with 20,000 Prices ................................................................. 94
3.27 Letters Report Meters .................................................... 94
3.28 Two-Factor ANOVA for Meters*Area ................................. 97
3.29 Letters Report Meters*Area ............................................. 97
3.30 Two-Factor ANOVA for Meters*Lower-Bound ....................... 99
3.31 Two-Factor ANOVA for Meters*Equipment-Price ................. 101
3.32 Letters Report Meters*Equipment-Price ............................ 102
3.33 Two-Factor ANOVA for Area*Lower-Bound .......................... 104
3.34 Two-Factor ANOVA for Area*Equipment-Price ..................... 105
3.35 Letters Report Area*Equipment-Price .............................. 106
3.36 Two-Factor ANOVA for Lower-Bound*Equipment-Price ........ 108
3.37 Sorted Parameter Estimates ............................................. 109
3.38 ANOVA Table ............................................................. 110
3.39 Summary of Fit Table .................................................... 110
3.40 Summary Results for Total Network Cost .......................... 114
3.41 ANOVA for Meters ....................................................... 116
3.42 Student’s t-test for Meters Comparing 5000 with 10,000 Meters .... 117
3.43 Letters Report Meters .................................................... 117
3.44 ANOVA for Area ........................................................... 119
3.45 Student’s t-test for Area Comparing 500 with 100 Areas ........ 120
3.46 Letters Report Area ....................................................... 120

xvi
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.47</td>
<td>ANOVA for Lower-Bound</td>
<td>122</td>
</tr>
<tr>
<td>3.48</td>
<td>Student’s <em>t</em>-test for Lower-Bound Comparing 200 with 20</td>
<td>123</td>
</tr>
<tr>
<td>3.49</td>
<td>Letters Report Lower-Bound</td>
<td>123</td>
</tr>
<tr>
<td>3.50</td>
<td>ANOVA for Equipment-Price</td>
<td>125</td>
</tr>
<tr>
<td>3.51</td>
<td>Student’s <em>t</em>-test for Equipment-Price Comparing 200000 with 20000 Prices</td>
<td>126</td>
</tr>
<tr>
<td>3.52</td>
<td>Letters Report Equipment-Price</td>
<td>126</td>
</tr>
<tr>
<td>3.53</td>
<td>Two-Factor ANOVA for Meters*Area</td>
<td>128</td>
</tr>
<tr>
<td>3.54</td>
<td>Two-Factor ANOVA for Meters*Lower-Bound</td>
<td>130</td>
</tr>
<tr>
<td>3.55</td>
<td>Two-Factor ANOVA for Meters*Equipment-Price</td>
<td>131</td>
</tr>
<tr>
<td>3.56</td>
<td>Two-Factor ANOVA for Area*Lower-Bound</td>
<td>131</td>
</tr>
<tr>
<td>3.57</td>
<td>Two-Factor ANOVA for Area*Equipment-Price</td>
<td>133</td>
</tr>
<tr>
<td>3.58</td>
<td>Two-Factor ANOVA for Lower-Bound*Equipment-Price</td>
<td>135</td>
</tr>
<tr>
<td>3.59</td>
<td>Sorted Parameter Estimates Equipment Count</td>
<td>137</td>
</tr>
<tr>
<td>3.60</td>
<td>Analysis of Variance</td>
<td>138</td>
</tr>
<tr>
<td>3.61</td>
<td>Summary of Fit</td>
<td>138</td>
</tr>
<tr>
<td>3.62</td>
<td>Summary Results for Equipment Count</td>
<td>141</td>
</tr>
<tr>
<td>3.63</td>
<td>ANOVA for Meters</td>
<td>141</td>
</tr>
<tr>
<td>3.64</td>
<td>Student <em>t</em>-test Comparing 10000 with 5000 Meters for Average Link Distance</td>
<td>145</td>
</tr>
<tr>
<td>3.65</td>
<td>Letters Report Meters</td>
<td>145</td>
</tr>
<tr>
<td>3.66</td>
<td>ANOVA for Area</td>
<td>147</td>
</tr>
<tr>
<td>3.67</td>
<td>Student’s <em>t</em>-test for Area Comparing 500 with 100 Areas</td>
<td>148</td>
</tr>
<tr>
<td>3.68</td>
<td>Letters Report Area</td>
<td>148</td>
</tr>
<tr>
<td>3.69</td>
<td>ANOVA for Lower-Bound</td>
<td>150</td>
</tr>
</tbody>
</table>
3.70 Student’s t-test for Lower-Bound Comparing 200 with 20 .......... 150
3.71 Letters Report Lower-Bound ........................................ 151
3.72 ANOVA for Equipment-Price ......................................... 153
3.73 Student’s t-test for Equipment-Price Comparing 200,000 with 20,000
    Prices ............................................................... 153
3.74 Letters Report Equipment-Price ...................................... 154
3.75 Two-Factor ANOVA for Meters*Area ................................ 156
3.76 Letters Report - Meters*Area for Average Link Distance ........ 156
3.77 Two-Factor ANOVA for Meters*Lower-Bound ....................... 158
3.78 Two-Factor ANOVA for Meters*Equipment-Price .................. 160
3.79 Two-Factor ANOVA for Area*Lower-Bound ......................... 161
3.80 Two-Factor ANOVA for Area*Equipment-Price .................... 163
3.81 Two-Factor ANOVA for Lower-Bound*Equipment-Price .......... 165
3.82 Analysis of Variance ................................................ 165
3.84 Sorted Parameter Estimates Average Link Distance ............. 166
3.83 Summary of Fit ..................................................... 167
3.85 Summary Results for Equipment Count ........................... 170
3.86 Summary of Experimental Observations and Hypothesis Testing Results 171

xviii
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>Advanced Metering Infrastructure</td>
</tr>
<tr>
<td>ALD</td>
<td>Average Link Distance</td>
</tr>
<tr>
<td>CLB</td>
<td>Conditional Lower Bound</td>
</tr>
<tr>
<td>DSM</td>
<td>Demand Side Management</td>
</tr>
<tr>
<td>DA</td>
<td>Distribution Automation</td>
</tr>
<tr>
<td>EC</td>
<td>Equipment Count</td>
</tr>
<tr>
<td>EP</td>
<td>Equipment Price</td>
</tr>
<tr>
<td>FoM</td>
<td>Figure of Merit</td>
</tr>
<tr>
<td>FCIFNP</td>
<td>Fixed-Charge Interval-Flow Network Problem</td>
</tr>
<tr>
<td>FCCMCNF</td>
<td>Fixed Charge Capacitated Minimum Cost Network Flow</td>
</tr>
<tr>
<td>FCIFNM</td>
<td>Fixed-charge Interval-flow Network Model</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>IFNET</td>
<td>Interval Flow Network</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>IIoT</td>
<td>Industrial Internet of Things</td>
</tr>
<tr>
<td>LB</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>MIP</td>
<td>Mixed Integer Linear Program</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>NPV</td>
<td>Net Present Value</td>
</tr>
<tr>
<td>NF</td>
<td>Network Flow</td>
</tr>
<tr>
<td>NOC</td>
<td>Network Operations Center</td>
</tr>
<tr>
<td>OR</td>
<td>Operations Research</td>
</tr>
<tr>
<td>PTMP</td>
<td>Point-to-Multipoint</td>
</tr>
<tr>
<td>PTP</td>
<td>Point-to-point</td>
</tr>
<tr>
<td>SG1</td>
<td>Smart Grid 1 model</td>
</tr>
<tr>
<td>SM</td>
<td>Smart-meter</td>
</tr>
<tr>
<td>SGCN</td>
<td>Smart-grid Communications Network</td>
</tr>
<tr>
<td>TNC</td>
<td>Total Network Cost</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
</tbody>
</table>
I dedicate this work to my family, Susan M. Klinkert, Lauren Sofia Klinkert, and Lawrence J. Klinkert, and parents Lawrence C. Klinkert, and Sofia P. Klinkert.
1.1. Introduction

This research addresses a critical area for utilities (electric, gas, water) of our country. These utilities can be viewed as networks that deliver resources from origination points to destinations at industrial and residential consumers. Improving the efficiency, security, and cost-effectiveness of such networks has been a national imperative for a decade, and the pace of improvements is accelerating. A utility network is typically accompanied by an electronic communications network that controls, monitors, and manages it. Both networks must and will improve significantly in the coming years to address the needs of effectively delivering the resource in a cost-effective, secure, reliable, and efficient manner.

The smart-grid is an advancement and automation of these two networks, the power systems network delivering energy, and the communications system providing command and control. The power system devices need communications and connection to databases within the utility and in some cases external to the utility (such as weather center services). As such, these smart-grid networks are the first widespread implementation of an Internet of Things (IoT) network. An IoT product, as defined by Sinclair [67], can be a physical device, system, or environment. For example an IoT product can be an individual smart watch, a system of smart meters, or an entire smart city. An IoT product contains smart devices, which can communicate, which can use data analytics and which can be connected to external systems and data, and bring to bear the full capability of the Internet [67]. An industrial IoT (IIoT)
network is typically a secure network with few, if any, direct connections to the open Internet.

The term used for improvements to these utility networks is “smart-grid” (SG), which implies that they will become more automated, or “smarter,” by embedding “intelligent” elements (microprocessors and microcomputers) in devices, then connecting them across a network. These devices, and the communications between them, perform critical functions such as reporting on the consumption in volume and demand in intensity of energy, gas, or water distribution; the quality of electric power; and the status of service, whether operational or failed. These intelligent devices must communicate with centralized and decentralized applications, analytics, and services in and across these new intelligent networks.

These new communications networks are termed smart-grid communications networks (SGCN). This study addresses the need of utilities to become more efficient in evolving these services and the SGCN is a critical and integral part of the electric, gas, or water utility of the future. This work provides an improved approach to the SGCN problem of designing a smart-grid communications network for utilities, in this work termed the SGCN (design) problem. This presentation is given in terms of an electrical power grid, although the approach is applicable to the distribution of other commodities, including water, natural gas, and other IIoT networks in general.

This work reviews the key literature published to date on the current and emerging ways of designing these new SGCNs or, more specifically, solving for an optimal SGCN topology. Then optimal topologies are generated and these topologies are solved for a range of example SGCN problems. Then an experiment is designed to understand the effect of key input factors on output figures of merit of the design process. The results are provided and will aid practitioners in planning, conducting, and evaluating these SGCN optimization problems.

This chapter delves into the problem context, reviews the relevant literature, presents an approach to solving the SGCN problem, and documents the expected
contributions of this work. As will be described in more detail, this work is focused on creating minimum-cost SGCN designs. Some work has been done in this emerging area, but little has been published specifically addressed by this problem.

1.2. Power-Engineering Context

To understand the SGCN design problem it is helpful to understand the structure of the past and present electric power delivery systems. It is also important to understand the future of power delivery, termed smart-grid, both the geographic expanse as well as the breadth of services or applications that comprise smart-grid.

To fully understand the importance of the communications network design problem addressed herein, it is key to appreciate the needs of the overall power systems structure for delivering electric energy. Figure 1.1 depicts the major activities associated with generating and delivering electric power to consumers.
The primary path of power flow is shown by the dotted lines across the bottom of the figure. This is the path of power flow from generation through transmission through distribution to the customer. The various clouds depict important areas of activity and the solid lines depict communications connectivity required between them to ensure the electric power system functions effectively, efficiently, reliably, and securely. The Network Operations Center (NOC), sometimes called the Data Center (DC), controls the operations of the automated and manual tasks associated with the reliable flow of energy through the electric grid. Other key stakeholders include the service provider, which acts as the “store front” to utility customers, and the energy markets that negotiate energy flows across generation regions.

Figure 1.1’s solid lines represent the communications links between all the stakeholders and domains needed to ensure the electric power system functions properly. This communications grid is critical to the functioning of the electrical power grid. The current structure for the power and communications networks is decades old, but modernizing. The question next becomes: How do we make these networks better? How do we make them “smarter?”

The term “smart-grid” also describes an increasingly automated electric system useful for emerging societal needs. For a utility, an automated system meets operations needs and increasingly important societal concerns. It is automated such that routine processes, events and work-flows occur with minimal human intervention. A SG meets societal expectations in terms of energy and resource efficiency, effectiveness, reliability, and security. Additionally, the SG should provide these services at a range of costs for consumers from basic, which supports a minimal but adequate level of service, to advanced services for those willing to purchase them. To date, electric utilities, versus other utilities (such as gas or water), have currently been the utilities that have deployed the largest number and types of SG in some parts of their jurisdictions. They still must, however, deploy other significant portions of SG in other parts of their full service territories.
In order for utilities to increase efficiency and effectiveness of their electric grid, they must make the system more automated and “self-healing.” A self-healing power system is one that can automatically detect that it is not operating in its nominally correct condition, and make any needed adjustments to restore the system to its normal operations. For large, complex and dynamic systems, such as the electric grid, the operators of the system strive to design-in a self-healing capability to automatically control various aspects of the grid with as little human intervention as is safely feasible. This automatic control typically originates from the central data center, but may also occur via automated machine-to-machine approaches. This automation goal is approached by overlaying a more capable communications network on the electric grid devices than has been used previously. Thus, at a high level, the SG is an increasingly automated and self-healing network, and a SGCN is needed to enable these operational and societal benefits.

A power utility typically operates over a wide area and an approach to realizing a SGCN must address the geographic expanse of a typical network. Figure 1.2 portrays the magnitude of the typical SGCN problem. In a utility’s geographic jurisdiction, the utility must service clients that are spread out over urban, suburban, rural, and remote areas. Customers in urban areas are usually densely concentrated, particularly in apartment complexes, business offices, high rise buildings. Communications to these areas are challenged by the need to communicate reliably in a densely populated environment, which for wireless communications technologies means communicating in the presence of significant spectrum congestion and physical barriers. For other technologies such as copper, coax, and fiber, this means physical congestion of the transport media in trenches, cable raceways, and conduits. For remote areas, the challenge is typically one of the “reach” of the technologies. For example, for wireless technologies there exist distance and line-of-sight limitations that make connections to remote devices difficult. Similarly for wired technologies, the distances involved make delivery of communications difficult and costly. Thus one key dimension in the
context of planning a SGCN is the coverage area and type of geography over which a utility must operate.

A utility must deliver a significant breadth of services and applications to realize a SG. Figure 1.3 typifies the end-to-end layout of a SGCN from the applications that provide services to customers, through the communications links that connect customers to the services. Applications developed by utilities provide either a service to customers or aid in the management and operation or automation of the grid. These services or applications can include: customer information systems (CIS), meter data management systems (MDMS), Advanced Metering Infrastructure (AMI), Distribution Automation (DA), Energy Management System (EMS), Outage Management System (OMS), Distribution Management System (DMS), Demand Side
Management (DSM), Work Force Management (WFM), Network Management System (NMS), and Geographic Information System (GIS). The literature is replete with information on what these application do. All these applications generate data communications traffic among themselves or to and from customers or devices on the grid. This research does not focus on the applications themselves, rather on the cumulative traffic generated by these applications, and how to connect them to realize a cost-efficient topology for SGCN infrastructure.

Another aspect of the services and applications needed for smart-grid is the breadth of communications functionality required for proper operation. Table 1.1 lists typical smart-grid communications data traffic needs and functionality required, as noted by the Department of Energy. Bandwidth and data rates can range from a few bits per second (bps) to megabits per second (Mbps) to handle the range of services needed. The latency requirements can range from a few cycles of AC power (milliseconds) to a few minutes. Across a large market, this drives the aggregated traffic onto a core wide-area network (WAN) transmission circuit to potentially several gigabit per second (Gbps) speeds. High-speed, low-latency communications links
are needed for the applications that protect the power grid, such as control links for circuit breakers, but some applications, such as AMI devices, also termed “smart electric meters,” need only provide low-latency, low data rates to perform their function. These data communications must also be provided with varying degrees of reliability, security, and availability through the use of backup power sources. This broad range of communications needs and functionalities creates a challenge for designers of the SGCN.

Table 1.1: Smart-grid Functionalities and Communications Needs (Source: [23])

<table>
<thead>
<tr>
<th>Applications</th>
<th>Network Requirements</th>
<th>Bandwidth</th>
<th>Latency</th>
<th>Reliability</th>
<th>Security</th>
<th>Backup Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td></td>
<td>10-100 kbps/node 500 kbps for backhaul</td>
<td>2-15 sec</td>
<td>99-99.99%</td>
<td>High</td>
<td>Not necessary</td>
</tr>
<tr>
<td>Demand Response</td>
<td></td>
<td>14 kbps-100 kbps per node/device</td>
<td>500 ms - several minutes</td>
<td>99-99.99%</td>
<td>High</td>
<td>Not necessary</td>
</tr>
<tr>
<td>Wide Area Situational Awareness</td>
<td></td>
<td>600-1500 kbps</td>
<td>20 ms - 200 ms</td>
<td>99.99-99.9999%</td>
<td>High</td>
<td>24 hour supply</td>
</tr>
<tr>
<td>Distribution Energy Resources and Storage</td>
<td></td>
<td>9.6-56 kbps</td>
<td>20 ms - 15 sec</td>
<td>99-99.99%</td>
<td>High</td>
<td>1 hour</td>
</tr>
<tr>
<td>Electric Transportation</td>
<td></td>
<td>9.6-56 kbps, 100 kbps is a good target</td>
<td>2 sec - 5 min</td>
<td>99-99.99%</td>
<td>Relatively high</td>
<td>Not necessary</td>
</tr>
<tr>
<td>Distribution Grid Management</td>
<td></td>
<td>9.6-100 kbps</td>
<td>100 ms - 2 sec</td>
<td>99-99.999%</td>
<td>High</td>
<td>24 - 72 hours</td>
</tr>
</tbody>
</table>

This section describes the context within which the problem of this research exists. It discusses the structure of the past and present electric energy delivery system, and the requirements of the future smart-grid in terms of communications infrastructure. It identifies the scope and scale within which this research exists,
that is, the geographic extent of a SGCN, as well as the scale of bandwidths and capabilities needed by the technologies to function as an effective SGCN.

The take-away from this context is that the SGCN is a large, growing, sophisticated communications network with a substantial flow of critical communications traffic. Designing these large, important, and costly networks – efficiently through the use of advanced optimization techniques and tools – is the goal of this research.

1.3. The Problem

The problem this research addresses is one of designing a communications network for smart-grid networks. The full smart-grid problem includes connecting all manner of smart devices on the grid to each other and to centralized applications at the data center or NOC. Such a problem involves specifying communications links among devices and between these devices and the centralized applications that monitor and control them. There is a large variety of smart-grid configurations and instances that are possible, given the growing number of applications, geographic situations, and customer requirements that exist. One way to approach such a problem is to focus on a part of the problem that, if solved, can be expanded to include any instance of the larger problem. This research thus focuses on the problem of connecting an important type of smart-grid device, a smart-meter, to an AMI application at the data center. This problem is faced by, or will be faced by, the majority of electric utilities. The problem this work will address is therefore the design of a SGCN with a smart-meter communications focus.

In designing a SGCN for a smart-meter-focused problem, the designer must use techniques from several disciplines. First a basic knowledge is needed of the commodity being transported, in this case electric power. Since these projects are typically large, involving many thousands of meters operating over many months or years, principles from engineering management and economic analysis are useful. Then a knowledge of the principles of electrical engineering is needed, specifically data com-
munications engineering. Finally this problem can benefit from the application of principles from operations research, namely network flow programming, and newly developed heuristics and software implementation for efficient optimization of SGCN-type network problems. The next section surveys literature from these areas useful in addressing this problem.

1.4. Literature Review

This section reviews the literature related to smart-grid communications networks with a focus on developing an optimal SGCN design. To set the stage, literature from the field of engineering management is briefly referenced regarding the planning of large engineering projects. Next, the literature from electrical engineering related to communication network design for the relevant technologies emerging that support SGCNs is briefly surveyed. Then literature is suggested from the field of operations research that applies optimization mathematics to the planning and design of the types of communications networks that support smart-grid.

With this backdrop, the power-systems engineering literature is identified that directly relates to the design of smart-grid networks, both the power delivery networks and the communications networks. An overview is provided of the relatively sparse amount of literature available that specifically addresses smart-grid communications network topology optimization using operations research techniques. This literature review shows that there are few published solution approaches existing for the top-down, end-to-end mathematical optimization of SGCNs. A brief section on experimental design methodologies provides references to works that support the experimental design performed in this research.

1.4.1. Engineering Management Literature

In developing a strategy, plan, and design for a large engineering project, such as a SGCN, general guidance is provided by classical engineering management texts.
For large projects, the total cumulative program costs (capital and operating) are relatively minor in the beginning of a project but, as the project moves to development, test, and commercialization, the total costs increase significantly, according to a seminal text in engineering management by Shannon [64]. Since many important decisions made in the early phases of a project “lock-in” costs and constraints throughout the project life cycle [64, p. 251], it is important to take the time during these early phases to carefully plan and — as proposed here — quantitatively model the envisioned system, to seek an optimal strategy for the project. Thus, quantitative modeling in the early phases is a cost-effective way to perform business analysis that screens out less-attractive approaches in favor of more-attractive, optimized approaches to realizing a project.

Another key guiding principle for large projects is to use systems-engineering principles [14]. When designing large projects, such as large SGCNs, it is advisable to advance the design through phases. These phases can be labeled conceptual, preliminary, and detailed design. Conceptual design allows for characterization of requirements, studying feasibility, and advanced system planning. A preliminary design phase allows for functional analysis, preliminary allocation of functionality, system and subsystem analysis, and detailed specifications. A detailed design and development phase establishes the final design specifications, support functions, test support elements, and potentially an initial pilot trial or prototype. It is useful to advance a model of the system through these phases to incorporate increasing amounts of detail, which allows for increasing levels of accuracy in the modeling of the final design.

1.4.2. Electrical Engineering Literature

Once general guiding principles are understood for realizing a large engineering project, guidance can be taken from a review of the broad area of electrical engineering related to communications networks. For the high-capacity core and distribution
layers of the network, design approaches are provided by numerous seminal works. They provide techniques for improvements to network reliability as well as deployment considerations, such as in Ramaswami [57], protection schemes such as mesh topologies by Grover [32], and some optimization techniques and tools in Cahn [18]. For the medium-capacity distribution and access layers of the network, wireless technologies are typically added to wireline and fiber technologies to handle the diversity of needs at these levels.

Large portions of a SGCN are comprised of wireless systems. Wireless network design methods are provided in several classical works, such as Taub [69], Goldsmith [29], and Rappaport [58]. These texts introduce the key concepts of modulation and coding, point-to-point and point-to-multipoint configurations, wireless path loss, and spectrum-efficient cellular system-design techniques. Radio propagation over smooth earth as well as over rugged, hilly and mountainous terrain is covered by Parsons [55].

As the bandwidth requirement increases, issues associated with broadband wireless systems are addressed by Anderson [3], which informs practitioners about how wireless systems behave differently as the wireless transport channel becomes wideband rather than narrowband. The prediction techniques change for the specialized modulation techniques needed by these wider channels, and thus how these signals behave when traversing terrain, morphology, and man-made obstacles. Specific broadband point-to-multipoint (PTMP) schemes are covered in by Korowajczuk [47] and Abate [1], including LTE, WLAN and WiMAX. Also point-to-point (PTP) microwave design is addressed in the classic text by Lehpamer [48]. Needed frequently by practitioners is Kobb’s practical guide [46] addressing the important challenge of where to find the spectrum to operate these technologies. Tools are available, as found in Sharma’s works [66, 65], that analyze specialized networks such as telephone-switching networks. Lastly, when needing to procure a cost-effective system of communications equipment from a variety of vendors, there are methodologies for specifying system requirements in such a way that vendors can respond effectively with accurate pric-
ing. One such set of methodologies for the WiMAX broadband wireless technology is described by Klinkert [44].

The background provided by a review of these works allows for an assortment of ways to design and deploy large communications networks. This review of communications engineering works provides a technological basis for planning a SGCN. It is clear there is a broad array of technologies that can be used to realize one. The preceding works, with few exceptions, provide only a descriptive understanding of the alternative technologies available and how to assemble them into some form of a solution. They do not provide a prescriptive approach, methodology, or tool for the design of a communications network, given a specific set of circumstances. The next set of resources do provide prescriptive approaches and tools for practitioners designing networks.

1.4.3. Operations Research Literature

The field of operations research (OR) provides a starting point for a prescriptive methodology for solving the SGCN design problem. This discipline provides, through mathematics optimization the best possible solution for a particular set of circumstances (given available inputs and needed outputs). OR provides prescriptive (as opposed to descriptive) solution guidance. This is useful especially in the early stages of project realization to make the strategy and planning phases more quantitative, and to base it on specific data values and a specified objective. The approach focuses on creating a quantitative strategy, plan, or design model early enough to make key decisions, before firm design specifications are finalized. When an OR model is solved, it prescribes a solution that a practitioner can consider.

The classic OR text by Hillier and Lieberman [34] introduces linear optimization models, as well as network and integer programming: the techniques needed to solve for ideal network topologies. A good grounding in graph theory by Chartrand [19] and linear algebra by Anton [4] provides the mathematics needed to precisely define
these mathematical models. For networks, Barr, Glover and Klingman [11] and Kennington [40] describes how to associate a fixed cost with establishing each link in the network as well as a variable cost ascribed to the variable data traffic flow on each link of an optimized topology. Another influential work by Ahuja, Magnanti, and Orlin [2] focuses exclusively on the theory behind the mathematical programming techniques for optimizing networks. Lastly, a pivotal work by Pioro [56] provides optimization models for modern Internet Protocol (IP) networks, multi-layer networks, and network topologies for reliability and resiliency. A recent important work by Kennington, Olinick and Rajan [41] applies OR techniques to wireless networks, including large-scale networks, of which a SGCN is one. These techniques are all prescriptive, in that they go beyond describing what technology is available, and how to design a network in general, to providing models and solution methods for identifying the best solution for a specific set of circumstances. These methodologies can be applied at all phases — strategy, planning, design, and implementation — of realizing a SGCN, thereby increasing the objectivity of each phase and, importantly, allowing modeling and “what if” analysis in the early phases of realization, when projects costs are lower, and significant constraining decisions have yet to be made.

1.4.4. Power-Systems Engineering and SGCN Design Literature

With this backdrop of knowledge of communications engineering and operations research methodologies for designing networks, the power-systems engineering area is reviewed for literature related to the design of electric utility and smart-grid networks. By understanding the underlying power systems network, one can more fully understand the overlaying communications network needed for its control. The vastness, importance, and criticality of the underlying power systems grid requires the overlaying communications system to be expansive, significant, and reliable.

The classic text by Glover, et al. [28] describes the basics of electric power systems including the fundamentals of voltage, current, power, phases, and power quality,
which are the “commodities” we wish to deliver reliably and cost effectively to customers. It also describes the power system network and power flows as well as stability, faults, and means of protecting the power grid from faults and transients. Emerging smart-grids are covered in the seminal work by the National Institute of Standards and Technology (NIST) [53]. In this classic, NIST provides a view of the challenges faced when society strives to meet widespread regulatory, societal, and technological drivers such as energy efficiency. Other works describe simply what smart-grid is (Department of Energy [22]), the governmental policies driving it (Kaplan [39]), and the future of smart-grid (MIT [51]). Other books by Fox [25] and Godstein [30] describe how new pricing pressures, energy conservation, environmental and green-planet initiatives play key roles in driving service and application-domain requirements for SG. Gellings and Books [26, 15] begin to merge engineering and smart-grid technology requirements as they discuss, in a descriptive manner, the approach to realizing an intelligent smart-grid network, including the standards evolving towards designing these networks. A thorough and important contribution to designing a SGCN by Budka [17] helps practitioners understand all the facets associated with designing these systems but, again, only in a descriptive manner, with only a passing reference [17, p. 172] to prescriptive approaches.

The smart-grid is but one early, large type of Internet of Things (IoT) network. Because it has emerged from utilities, it is also termed an Industrial Internet of Things network. Discrete products (like a single unique smart grid sensor), a system (like the utility’s entire smart-grid system) or an environment (such as a smart city) become IoT products when they are associated with external systems on the larger Internet [67]. While security and privacy remain paramount, smart-grid systems today are increasingly connected to external Internet systems. External systems that utilities connect to include weather, industrial control systems, and smart home systems. These connections will only grow, leading to the term “Internet of Energy” and “connected objects” [33]. Wide-area IoT systems and environments require a
communications network in many ways similar to a smart-grid communications network. This work, therefore, can be extended to support the design of wide area IoT networks. Many papers have recently been published that address smart-grid as an IoT system [67, 33, 72, 71, 59, 42, 21, 50].

New, large, and complex smart-grid power systems require a similarly complex SGCN for its command and control. The OR literature is rich with techniques to do so. Practitioners can gain experience with network solutions early in the process, as they plan, design and deploy these networks.

1.4.5. SGCN Design Literature

A recent literature survey of SGCN communications network topology optimization—central to the research focus of the present work to solve the SGCN problem—discovered the following studies. In a paper by Jahromi [35], the authors provide a (non-linear) genetic algorithm and model for the solution of an optimal SGCN. A work by Li [49] developed an approach to solving the SGCN problem by means of clustering algorithms.

As shown above, while there is abundant information on communication networks for the typical access, distribution and core layers of communications networks, and there is much new work depicting and describing SGCNs, there appears to be works in the literature focused on a methodology for deriving prescriptive topology optimization of SGCNs, even in the OR or SGCN-focused literature. There are helpful resources by Ahuja and Pioro [2, 56] on network flows for communications networks and these can be the grounds for developing network topology optimization for SGCNs, as is done in this study. Several recent papers approach a subset of the challenge, but no paper addresses the subject by focusing on a holistic, top down, end-to-end modeling and solution approach. This work will focus on providing a methodology to define requirements and solve for an optimal SGCN topology, using a suburban smart-meter scenario.
1.4.6. Experimental Design Literature

In this section, the statistical sciences are reviewed for the relevant experimental techniques useful for statistical modeling of the results of the testing conducted for this work. These sciences are reviewed for techniques practitioners can use to conduct statistical confirmation of optimization work performed, as done in the experimentation portion of this document.

The classic text by Montgomery [52] for design and analysis of experiments provides the fundamentals in designing the experiment described in this research. It documents the techniques for leading to objective conclusions by use of hypothesis testing, as well as regression analysis. Hypothesis testing allows for the testing and selection of one hypothesis over another through the use of statistical techniques. Regression analysis assesses the degree of association between selected factors and experiment response variables. It also addresses the important concepts of coded experiments, orthogonal experiments, and optimally designed experiments, all relevant to the present work. Additionally, a second more practical cookbook approach to industrial experiments is the reference by Schmidt [63]. On-line works are emerging, such as by Parris [54] and Balka [8], that clarify many of the techniques in the experimental design process applied in the present effort.

This section reviewed of the literature related to smart-grid communications networks and explored published work that quantify strategy, planning, and design of communications networks using not just descriptive, but prescriptive techniques, such as mathematical optimization. The emerging smart-grid literature to date provides the requirements for addressing the SGCN problem but requires integration of methodologies, tools, and techniques from engineering management, electrical engineering, and power systems engineering. Operations Research provides the prescriptive tools useful for modeling specific design problems and identifying an optimal design. The statistical sciences area provides techniques needed to statistically eval-
uate the results.

1.5. The Solution Approach

What follows is a description of a prescriptive approach to modeling and solving the problem of SGCN design. This is not widely done, presumably due to the fact that network flow programming is not well-known to practitioners in this area, and the IFNET solver used herein is a new tool available for these types of problems. This software, along with a specific modeling methodology is a solution approach that addresses the current industry practice of manual network design, which results in suboptimal solutions due to inefficiencies and inconsistencies of human judgment injected at all stages and layers of a large SGCN program. A new advanced solver is introduced that addresses the need for fast solutions with promise to solve large-scope and large-scale network optimizations needed by designers for the SGCN problem.

The following is the overarching methodology within which this solution approach resides. It is the proposed methodology for solving for the optimal SGCN topology, using the mathematical model formulation to be described in the following section.

Underpinning this methodology is the availability of a new advanced algorithm and code for solving the SGCN problem. Barr and Apte developed the first interval-flow algorithm and code using a network code by Barr [5] and subsequently advanced by Barr and Jones to create the solver code named IFNET [37]. Using the interval-flow technique to identify solutions with the IFNET code, a SGCN model is formulated and used here to solve various model instances. Given this advance in technology, realization of a cost-effective SGCN is now more readily achievable. Using this advanced capability, the following approach is offered.
Methodology for solving the SGCN problem

1. Objective: Set the objective and guiding principles for the current phase (strategy, planning, design, or deployment);

2. Data: Collect the relevant infrastructure data (traffic sources, traffic destinations, equipment options, costs and candidate sites for network equipment placement);

3. Mathematical Model: Mathematically model an abstraction of the problem (the SGCN Model in Section 2.4);

4. Preparation: Specify scenarios and assumptions, generate problem instances, and with appropriate optimization software, solve test problems and tune parameters.

5. Optimization: Perform multiple optimizations on a range of instances to determine the optimal (minimum cost) topology;

6. Ideal Architecture: Use the optimal topology as the governing ideal conceptual architecture for each major area of the network (core, distribution, access) and finalize second- and third-order details through the use of specialized deep design tools and techniques;

7. Lead and Manage: Lead the realization process of the SGCN throughout the planning, design, implementation, and operations stages, re-optimizing as necessary, as each stage is reached.

This approach is best used at the start of a system’s strategy phase and updated across subsequent planning, design, implementation, and operations phases. However, it can
be invoked at any phase that seeks guidance from an ideal topology while designing, deploying, or operating a large and complex network.

This approach creates the optimal SGCN topology and concomitant SGCN architecture as early as the strategy phase. It can be used to update and refine the incumbent ideal topology for continued use in all subsequent phases.

The above methodology requires fast solutions for large-scale SGCN problem instances. To do this, the recent advance in network flow programming software termed IFNET, has been realized at Southern Methodist University. The genesis of the present-day solver was a paper published in 1979 of a fast algorithm for pure network flow problems and implemented in a code called Arc-II [9], later Netstar. Netstar has been modified for other problems since, such as fixed-charge network problems [11], and in parallel processing applications [12]. In 1998, Barr proposed an interval-flow model [6] for certain types of network models and, along with his student Apte, modified Netstar to implement a new interval-pivoting heuristic that quickly solved these mixed-integer problems. Jones [7] in 2004 used the interval-flow Netstar after re-implementing it in C to solve a broader class of interval-flow networks. The code is named IFNET, and is an advanced solver for interval-flow networks. This solver combined the numerous advanced algorithms into an overarching heuristic approach that quickly finds feasible and minimum cost solutions to the computationally challenging problem class.

The effectiveness of the solver was tested against a state-of-the-art commercial tool and the results are published in a separate work by Barr and Jones [37]. In experiments performed in that work, IFNET was shown to be over an order of magnitude faster than the top-tier commercial solver when solving general network topology problems. Also, in a recent unpublished work, Barr, Jones, and Klinkert [13] demonstrate that the IFNET software maintains the order of magnitude speed and high quality of solutions for various instances of typical smart-grid type communications networks. The results of these experiments have proven that high-quality solutions
are attained, and provide important decision-support prescriptions, with more than an order of magnitude increase in speed even when used on a basic personal computer. This bodes well for solving much larger networks with this technique through the use of well-known and emerging cloud-based parallel processing, graphical processing unit (GPU) power, and other code size and speed optimization techniques. Thus, this advance in solving the considered SGCN topology problem provides a key element in the proposed methodology as a solution approach to the problems with current realization of SGCNs.

In summary, this section presented a methodology based on a new advanced heuristic and solver that ameliorates problems with subjective human judgment in the realization of SGCNs. By using a governing optimal topology as the underpinning of a governing architecture, and by consistently using this as the ideal model in realizing SGCNs, designers of SGCNs are guided by prescribed minimum cost architectures throughout the realization of these networks. The methodology listed here provides a standardized approach to the realization of a SGCN and reduces instances of suboptimal strategies based on subjective human judgment. Such a model provides the ability to solve for the most important “first order” factors of traffic and connectivity for proper dimensioning of the entire network in terms of capacity, utilization, and topology at all network layers. The methodology provides the ability to quickly obtain automated solutions, gracefully scalable in terms of size as the network grows, toward providing key decision support to engineering management.

1.6. Expected Contributions

Through the use of the proposed methodology, the use of a breakthrough heuristic, as realized in an advanced new solver, this work is expected to contribute to engineering management’s ability to effectively realize SGCNs in a cost-controlled manner. By developing an accurate model of the SGCN problem focused on a smart-meter
deployment, and applying an advanced computing heuristic for the optimization of problems of this type, this research delivers a methodology for quickly and accurately solving for the minimum cost topology for a SGCN.

The experimentation conducted as part of this research supports solution instances of moderate size, although it scales to large size. By providing a solution to the SGCN problem, this research contributes to the science of developing strategy, planning, designing, and deploying efficient smart-grid programs, thereby enabling the efficient and effective distribution of energy for societal needs.
Chapter 2

PROBLEM DEFINITION AND MODELLING APPROACH

This chapter addresses the SGCN design problem, its formulation as an optimization model, and a set of problem instances for experimentation. First, the research problem is carefully defined. Next a series of simple to more-advanced models are formulated that support SGCN design. Finally, several instances of a complete model are set forth for use in subsequent chapters for experimentation.

2.1. Problem Definition

In high-level terms, an important smart-meter communications-network problem is: How should all smart-meters at customers’ homes be connected for communication with the network’s data center or operations center? Key to answering this question for a large network is to mathematically model and optimize the topology of the network’s access, distribution, and core elements that constitute a multi-level, hierarchical SGCN. The problem addressed here is the design of a smart-meter network for an electrical-power system, a design that establishes the selection, placement, connections, and connection capacity of the equipment required to meet a given set of demands at minimum cost.

This problem can also be stated as: “What is the lowest-cost topology for a given smart-grid communications network?” The answer requires finding the network topology with the lowest-cost set of communications links and equipment with suitable capacity to support all communications-data traffic sources with one or more traffic destinations. The resulting topology should also address the question: “How do I connect all data traffic sources to one or more traffic destinations in the most
cost-effective manner?” These notions are the subject of this research, and further developed in this chapter.

In setting out to solve this SGCN problem, a series of mathematical models are formulated, starting with a straight-forward version and progressing step-wise to one that encompasses the realism encountered in practice. To demonstrate the capabilities of the full-featured model, a collection of example problem instances are defined for subsequent experimentation.

2.2. Mathematical Notation, Conventions and Definitions

This section sets forth the mathematical notation used in the formal specification of model elements. For the formulation of the mathematical models that follow, the following notations and conventions are used.

Scalars are denoted by italicized lowercase Latin and Greek letters. Sets are denoted by italicized uppercase Latin and Greek letters. The set created by removing element \( e \) from set \( S \) is given as \( S - e \). A functional operator, the ordinal, is defined as \( ord(i,I) \) and is the ordinal sequence number of element \( i \) within an ordered set \( I \). Thus \( ord() \) can be considered an integer “position number” of the element within the set. These ordinals range in value from 1 to \( |I| \), the number of members of the set. An ordered subset of ordered set \( I \), consisting of the \( i^{th} \) through the \( j^{th} \) elements, is given as \( \Upsilon(I,i,j) \). The number of elements \( n \) in a set is denoted by \( ||N|| \), the cardinality of the set. Thus, for the set \( N = \{1,2,3\} \), \( ||N|| = 3 \).

The above descriptive terms and mathematical notation are used in the following descriptions of models for the careful specification of mathematical statements needed to precisely state a model formulation. The consistency of the mathematical notation is critical for the correct specification of a mathematical model. However, for clarity and some practitioners, the terms sites, locations, arcs, nodes, and traffic shall be defined and may be used when the context requires application-specific terminology.
This is important in understanding SGCNs, their topologies and architectures.

2.3. Formulation of a SGCN Design Optimization Model

This section establishes the formulations leading up to a mathematical model for the problem of designing a SGCN. These models are, broadly speaking, known in the field of operations research as mathematical programming models.

A mathematical programming model consists of (1) a set of decision variables, (2) an objective function and (3) a set of constraints. Decision variables are unknown quantities whose values are to be determined and under user control. Clear definitions of decision variables are critical to model formulation and interpretation. For the proper definition of a decision-variable [10], one should state the:

- **Units of Measure**.
- **Subject** (a noun), describing the what or who involved.
- **Action** (a verb) and what is happening to the subject.
- **Location** by describing where the action would occur.
- **Time** when the action would happen to the subject.

A typical variable definition would take the form: variable name = number of (Units of measure) of (Subject) to (Action) at (Location) during (Time). As an example, the flow of communications data traffic on a unidirectional communications arc can be defined as the variable $x$ which equals the traffic flow in units of Kbps, traversing over the connection from point $i$ to point $j$ during steady state conditions. Another example is a binary variable that can be defined to indicate whether data traffic flow exists on a arc or not. This variable can be defined as $y$ equals to an indicator variable taking on dimensionless unit values of 0 and 1 that serves that takes on a binary 1 or 0 value, on the arc, when flow exists or does not exist on the arc.
A set of assigned values for all model variables is called a *solution*. Since a given model may have an infinite number of solutions, there needs to be a way to compare them. This is the role of the model’s *objective function*, a mathematical expression involving the variables that is to be maximized or minimized. An example objective is to minimize total cost; the corresponding objective function would be a linear expression that calculates the cost of a solution. When a solution is substituted into the objective function, a single value results, called the *solution value*. Hence the values of different solutions can be compared to choose the most preferred one.

But not all solutions are appropriate for a model. *Constraints* are mathematical statements that express limits on variable values and combinations of variables. Linear constraints involve a linear expression, a constant, and a relational operator. The three possible relational operators are: equals (denoted by “=”), less-than-or-equal to (“≤”) and greater-than-or-equal-to (“≥”). These can be used to express solution restrictions, such as:

- The value of variable $x$ cannot exceed 100 ($x \leq 100$);
- The value of $x$ cannot be negative ($x \geq 0$);
- The amount spent cannot exceed the budget maximum ($10x + 20y \leq 1000$);
- The value of $x$ must be exactly twice the value of $y$ ($x = 2y$ or $x - 2y = 0$);
- A solution must have at least 10 units of $x$ and $y$ combined ($x + y \geq 10$).

A given solution will either violate or satisfy a constraint. When the values of the solution’s variables are substituted into a constraint, the stated relationship expressed is either true or false. If false, the constraint has been *violated*; if true, that constraint is said to be *satisfied*. Only the solutions that satisfy every model constraint are of interest; these are called *feasible solutions*. If a solution violates any model constraint, it is considered to be an *infeasible* solution.
The scalar constants used in defining a model’s constraints and the objective function are termed the model parameters. The parameter values are often referred to as the model “data,” since the mathematical model can be defined using any number of parameter sets or “model instances.”

A complete model, then, consists of (a) a set of properly defined decision variables, (b) a set of constraints, (c) an objective function to be optimized (maximized or minimized), and (d) a set of parameters. A model is solved by identifying its optimal feasible solution: a set of variable values that satisfy all constraints and have the best possible solution value. If a feasible solution cannot be found, the entire model is considered infeasible or unsolvable. A model is unbounded if a variable can be increased without limit, which usually indicates a formulation error.

These powerful mathematical models are widely used for designing and operating systems in modern society. Whether identifying the shortest route to your GPS destination, designing a modern aircraft, determining the best prices of product and service, or managing power generation throughout an electric grid, mathematical programming models are used to improve systems and everyday life worldwide.

Many problems in practice can be addressed by techniques of mathematical programming. Several classes of solution algorithms have been developed that depend on the requirements of the problem, and the constraints used to characterize the problem. The most straightforward class includes problems with equations and an objective function that are all linear in the variables, which take on real-valued numbers. Solution techniques for this class of problem is addressed by the area called linear programming. Linear programming has a readily available solution algorithm known as the simplex method [34, p. 109]. Problems that can be modeled this way have an effective solution method that is used in all state-of-the-art optimization software.

Problems that can only be modeled using a non-linear equation in the constraint(s) or objective function can be more challenging to solve. Solution techniques for this
class of problem are the focus of non-linear programming [34, p. 654]. Non-linear models can be computationally difficult to address and a variety of solution techniques and heuristics have been developed for them.

Another challenging category of optimization problems has constraints requiring variables to take on integer values. In these integer programming (IP) problems, feasible solutions cannot have decision variable values that have a fractional term. For exact solutions, these problems require special solution techniques, including branch-and-cut and branch-and-bound, which can exhibit exponentially long solution times. For challenging problems, heuristic techniques have been developed to provide high-quality solutions quickly [34, p. 576]. Closely related to these problems are those requiring a combination of both real-valued and integer-valued variables in a model, which is termed a mixed integer model and are the focus of mixed integer linear programming (MIP) [34, p. 576]. Again, these models can be difficult to solve and for which exact and heuristic methods have been developed. Lastly, in the broad and dynamic area of mathematical programming, many other problems with special requirements beyond the above are areas of active research for techniques and solution approaches.

One important set of problems mathematical programming can effectively address is that related to networks. According to Webster, a broad definition of a network consists of “an interconnected or interrelated chain, group, or system” [68]. In the context of communications systems, a more focused definition of a network is given by Green [31] as “a set of communications points connected by channels.” Operations research practitioners have created models for problems in transportation networks, social networks, and communications networks, among many others.

A special form of linear programming is called network flow programming, which is relevant to formulating models for designing SGCNs. These network models fortunately take on a special form that allows for efficient algorithmic techniques to be applied toward their solution. The problem variables and constraint equations that
define network connectivity have a pattern conducive to efficient solution methods. Next is a description of the elements necessary to formulate and solve network-flow models. Beginning with the simplest network-flow model.

Network flow models can be portrayed by a graph with a collection of objects called “nodes,” depicted as circles, and “directed arcs,” depicted as arrows between node pairs. When structured appropriately and supply and demand values added, these sets of nodes and arcs become a graphical representation of a network flow model. One can start as basic as possible by connecting two nodes with one arc, thereby forming the elemental building block of a network, as shown in Figure 2.1.

![Figure 2.1: Model of a Communications Arc](image)

This one-arc network element is characterized in detail as follows.

In the figure, $N = \{i,j\}$ is the set of nodes, $A = \{(i,j)\}$ is the set of directed arcs, $n = |N|$ is the total number of nodes and $a = |A|$ is the total number of arcs. In this case there are $n = 2$ nodes, labeled $i$ and $j$, and one arc, $(i,j)$. The nodes can represent communications network sites where equipment is placed. As such, a node is defined here as the site or the equipment at the site needed to establish one connection of a communications arc. An arc is defined here as a connection between two nodes that allows flow of a commodity between the node pair in the direction of the arrow.

The above has a supply of $b$ units of communications-data traffic. This scalar can
represent units of data traffic flow measured in bits per second (bps) (or Kbps, Mbps, etc.). The source of the traffic supplied to the network element is indicated by the arrow from $b$ to node $i$. The demand, which is the quantity of communications traffic to be removed from the network element from node $j$ to $b$. This network element has a demand equal to supply of $b$ units of communications data traffic, and has the same units. The traffic output from the network element at node $j$, or the demand from the network element, is therefore indicated by an arrow from node $j$ to $b$. Thus, the input to the network element is the supply or source traffic and the output of the network element is the demand or destination of the traffic.

In this elemental network we have $a = 1$ arcs, which represents possible flow of a commodity from an ingress node $i$ to an egress node $j$ over arc $(i,j)$. As mentioned, the commodity of flow is communications network data traffic, or simply traffic, defined in this example as the steady-state flow in bits-per-second (bps) from the ingress node to the egress node. The decision variable $x_{ij}$ is defined as the amount of flow of traffic on arc $(i,j)$. Note the arrow emphasizes that this traffic cannot flow in the reverse direction from $j$ to $i$. This is a graph of a unidirectional communications arc. A bidirectional arc would be modeled with a second arc $(j,i)$ for traffic flowing from $j$ back to $i$.

A node cannot by itself internally generate nor dissipate traffic, so the sum of flows into and out of a node must sum to zero. Thus a node is an exchange point for traffic. Since a node is an exchange point only for traffic-in to traffic-out, and cannot itself internally generate nor dissipate traffic, the flows into and out of a node conform to the law known as conservation of flow. In this case we have a flow into node $i$ of $b$ and, given conservation of flow, the flow out of node $i$ must be $b$. Thus the flow on the arc must equal $b$, as well as out of node $j$. The total traffic $b$ entering this network also exits the network as $b$.

The parameters $l_{ij}$ and $u_{ij}$ represent, respectively, the lower and upper bounds on the flow on the arc, such that $x_{ij} \geq l_{ij}$ and $x_{ij} \leq u_{ij}$. These bounds constrain the
traffic flow on the arc to be between these levels. This provides the ability to model equipment with a capacity limit, or an arc that must flow a minimum amount of traffic. The arc can only flow traffic between the lower and upper bound constraints.

Given a graph of a model, and the mathematical model itself, the key to solving a general model is to find the value of the decision variable \( x_i \), the flow on the arcs, such that the objective function is optimized (minimized in this case) and satisfies all constraints. The parameter \( c_{ij} \) is the cost per flow unit of traffic, for example $1 per gigabyte of data traffic flow per second. The parameter \( f_{ij} \) is a one-time cost if the arc carries flow, also termed the fixed charge of the arc. It includes the cost contained in both nodes allocated to establishing the arc.

With these simple definitions and conventions in mind, one can model much larger networks, and use mathematical techniques to solve for various solution objectives, such as solving for a minimum cost network topology, given capacity constraints on the arcs. Large networks could have \(|N|\) nodes and \(|A|\) arcs in the millions.

For larger networks, one can model the possible connections between nodes, bounds on traffic flow on the arc, the cost of constructing the arc, and the cost of flowing traffic on the arc. Given an input amount of flows one can then use the mathematical model of the network to solve for the amount of flow on all arcs, and the solution for all decision variables \( x_{ij} \). Along with this, a solution value is obtained that includes the resulting total cost for the network. This is powerful and valuable when scaled to large networks.

With the above in mind, it is instructive to work through an example of a simple problem, model formulation, and solution. Take for this example what is known as a transportation network flow problem. The graph of the model of an example transportation problem is shown in Figure 2.2. In this simple transportation network, the nodes can represent warehouses and the arcs can represent highways connecting the warehouses. Nodes in this example can act either as a source, supply, or input to the network or as a destination, demand, or output from the network. Thus goods
must flow from supply warehouses $a$ and $b$ to destination warehouses $c$, $d$ and $e$. For the arc $(a,c)$ from node $a$ to node $c$, there is a decision variable $x_{ac}$ that represents the quantity of goods to flow over that arc. There is also a per-unit cost of flow, $c_{ac}$ on that arc of 10 times the quantity of goods transferred over that arc. For this simple problem, there is no fixed charge ($f_{ac}$), lower bound ($l_{ac}$), or upper bound ($u_{ac}$). All other arcs have similar decision variables and unit cost parameters for the flow on those arcs (for clarity, variables and parameter names are not shown). The supply consists of a total of 300 units of goods to flow into the network, and a total demand
of 300 units of flow out of the network. Lastly note this simple example does not have any upper or lower bounds, nor fixed charges on arcs, but still conveys the general process of formulating network flow models.

In terms of solutions to this problem, there can be many feasible solutions. Feasible solutions in this example are ones in which flows on all arcs meet the constraints of: (1) flow only on and in the direction of some or all arcs, and (2) the requirements of the supply (source flow) and the demand (destination flow). An optimal solution is a minimum-cost feasible solution (there may be more than one in certain special cases).

In order to solve the simple graph depicted in Figure 2.2 and identify the optimal solution, one must formulate mathematical statements and structure them into a solvable set of objective function and constraint equations. The objective function is written so as to minimize the total network cost by containing cost parameters multiplying decision variables containing the amount of flow in each arc. The cost is minimized in this case by solving for the flow in each arc. The constraint equations in this example are those that constrain the flow to only the arcs shown and directions shown. These constraint equations are equations written for each node to preserve the conservation of flow. Thus the flow-balance equation at each node requires that the total flow into a node equals the total flow leaving the node. Lastly, there are to be no negative flow of goods. This requires a last set of equations known as non-negativity constraints, here combined into one statement. Therefore, for the simple problem in Figure 2.2, the following is the structure of the mathematical statements of the objective function and, subject to, the constraint equations, gives the complete mathematical programming formulation for this problem.
Minimize

\[ z = 10x_{AC} + 20x_{AD} + 30x_{AE} + 20x_{BC} + 10x_{BD} + 5x_{BE} \]

subject to:

\[ x_{AC} + x_{AD} + x_{AE} = 100 \]
\[ x_{BC} + x_{BD} + x_{BE} = 200 \]
\[ x_{AC} + x_{BC} = 100 \]
\[ x_{AD} + x_{BD} = 50 \]
\[ x_{AE} + x_{BE} = 150 \]
\[ x_{AC}, x_{AD}, x_{AE}, x_{BC}, x_{BD}, x_{BE} \geq 0. \]

Solving this linear model using linear programming techniques [34] yields an optimal solution of:

- \( x_{ac} = 100 \)
- \( x_{ad} = 0 \)
- \( x_{ae} = 0 \)
- \( x_{bc} = 0 \)
- \( x_{bd} = 50 \)
- \( x_{be} = 100 \)

And solution value (total cost of the network)

\[ z = 3250 \]

From a practitioner’s point of view, the solution of this problem is helpful. By solving this problem, it becomes immediately clear what arcs are not needed (\( x_{ab}, x_{ad} \) and \( x_{ac} \)), what arcs are needed (\( x_{ac}, x_{bd} \) and \( x_{be} \)) and what the precise amount of flow is on each of these arcs (\( x_{ac} = 100, x_{bd} = 50 \) and \( x_{be} = 100 \)). Due to the cost of flow on each arc, the total cost of the network can be found to be \( z \).

The above is an example of a pure network problem, which has the property that if the supplies and demands are integer-valued, the flows will be integer-valued. This particular problem is in the form of a transportation problem such that “goods” are to be routed from supply to demand nodes. This problem is termed a bipartite problem since arcs only connect nodes with supplies to nodes with demands and there are no intermediate nodes.
2.3.1. Uncapacitated Minimum Cost Network Flow Model

Given the definitions and examples above, a general mathematical model is now described for pure network problems. The model is applicable to the simple transportation network of Figure 2.2. This first basic network model formulation is called the uncapacitated minimum cost network flow (MCNF) linear programming model. In this case, flow on arcs are not bounded by a capacity limit, thus the term “uncapacitated.” This highlights that no flow-capacity limits (upper bounds) are placed on any of the arcs. This directed network \( G(N,A) \) is mathematically modeled as a set of nodes \( N \), and a set of arcs \( A \) such that each arc connects a node \( i \in N \) to a node \( j \in N \), stated mathematically as \( A = \{(i,j) | i,j \in N\} \). For this model graph, one can solve for the flows on the arcs by applying methods known as network flow programming (as described by Ford [24], Kennington[40], and Ahuja[2]), which provides specialized techniques for identifying an optimal solution. Instances of large networks have been solved by these techniques (Barr [12]).

The mathematical model for an uncapacitated MCNF is as follows. The optimization problem of finding the minimum-cost set of arc flows to route all data through the network is defined by the following mathematical statement:

**MCNF:**

\[
\text{min} \sum_{(i,j) \in A} c_{ij}x_{ij} = z
\]  

subject to:

\[
\sum_{(k,j) \in A} x_{kj} - \sum_{(i,k) \in A} x_{ik} = b_k, \forall k \in N
\]  

\[
x_{ij} \geq 0, \forall (i,j) \in A
\]
where

\( z \) is the total network cost to be minimized

\( x_{ij} \) represents the amount of flow on directed arc \((i,j) \in A\) from originating node \(i \in N\) to destination node \(j \in N\),

\( c_{ij} \) is the unit cost of flow on arc \((i,j) \in A\), and

\( b_k \) is a scalar requirement of flow units at node \(k \in N\), positive if a supply, and negative if a demand.

This model is a linear program, from which one can determine the flow on all arcs. Practically, this amounts to the flow on the set of arcs required to realize a network that is not only feasible but optimal in terms of cost (minimum cost) or other linear objective. Some arcs may have zero flow. The optimal solution is therefore the resultant network of flows on arcs. The solution value \( z \) is the total cost of the network, which is to be minimized. This total cost is comprised of the cost of the flow of traffic on all of the arcs.

Figure 2.3 shows another application of MCNF in which there exist intermediate transshipment nodes that can serve as intermediate routing points for the transport of physical goods or communications data traffic from input to output nodes. A transshipment node would be used if it enabled an overall lower cost can be achieved for the network, for example when aggregation or distribution of flow might make the network more cost efficient. Transshipment problems allow the incorporation of candidate locations or available sites, which then allows for the identification of optimal intermediate distribution/aggregation points [34].

In a practical communications network context, communications sites are modeled as one or more nodes, and communications arcs are modeled as one or more arcs. A bi-directional link is modeled as two directed arcs connecting the same node pair in
opposite directions. Some sites in a problem are required in the network, such as customer homes and businesses to be connected into the network. These required sites typically have meter data traffic modeled as inputs to the network. Also, some required sites will be given as traffic output sites, such as a data center(s) or network operations center(s) where all meter traffic is received and taken out. Some sites are not required, rather are candidate sites or available sites that are potentially useful to aggregate traffic in the upstream direction or distribute traffic in the downstream direction. These candidate sites can be used to aggregate flow, but also to overcome distance limitations. For example, most technologies have range (distance) limitations, which can be modeled as arcs reaching nodes within a certain distance. In such cases, candidate sites can be used to bridge a longer distance through a series of “hops” through intermediate nodes.

This type of problem has inputs nodes, transshipment (or for communications networks, candidate or available) nodes, and output nodes. Available sites them-
selves can be a source or destination of network traffic. Thus, applying this classical transshipment problem to the SGCN problem, one can model the practical situation of having traffic flows from input sites of individual meters, to available-to-use intermediary sites (such as utility poles, electric substations, or radio towers), to output sites. This achieves a minimum cost network topology that transports traffic from meters to data centers.

2.3.2. Capacitated Minimum Cost Network Flow Model

The above models allow any amount of traffic on each arc. This would apply if communications equipment could connect sites (nodes) with any amount of traffic flow on communications arcs. This is not the case for many networks, since there are limits to the amount of traffic that communications equipment can transport. In order to eliminate that limitation in modeling SGCNs, the use of a “capacitated” network flow model is required. Thus, an arc \((i,j)\) can have a flow, \(x_{ij}\), and unit cost, \(c_{ij}\), but also an arc capacity limit or upper bound, \(u_{ij}\). These bounds are constraints to limit the maximum arc flow \((x_{ij} \leq u_{ij})\).

In order to model Figure 2.3 with capacitated arcs, the mathematical model for MCNF is now modified as follows.

\[
\text{MCNF/C:}
\]

\[
\min \sum_{(i,j) \in A} c_{ij}x_{ij} = z \quad (2.5)
\]

such that:

\[
\sum_{(k,j) \in A} x_{kj} - \sum_{(i,k) \in A} x_{ik} = b_k, \forall k \in N \quad (2.6)
\]
\[ 0 \leq x_{ij} \leq u_{ij}, \forall (i,j) \in A \quad (2.7) \]

where \( u_{ij} \) is the upper bound on the flow on arc \((i,j)\).

### 2.3.3. Fixed Charge Capacitated Minimum Cost Network Flow Model

While MCNF/C is useful for modeling a large class of industrial problems, some problems have additional costs that it does not incorporate. What is lacking for a SGCN model is the cost of initially constructing the arcs of the network (i.e., the equipment at each node required to establish the node-to-node connectivity). This initial “first cost,” “fixed cost,” or “fixed charge” required to construct an arc is classically termed the “fixed charge” of an arc in the network and represents the initial one-time cost associated with implementing an arc. In solving for the total cost of a new or expanded network topology, both the cost of establishing the communications arcs and the cost of the flow on arcs are needed. Thus, both the variable cost and fixed charges determine the total cost of each established and operating connection. The cost of establishing each arc can be viewed as the cost of the equipment in the nodes on both ends of each connection and the installation of any physical connections in between. More generally, the fixed charge can be a lumped cost of any or all of the following: the cost of the equipment on both ends of the one-way arc (half the cost of a two-way link), the cost of the site and site infrastructure allocated to the arc (such as the implementation costs), and a present-value lump-sum of the ongoing operations and maintenance costs allocated to the arc. A fixed charge model, termed a fixed charge capacitated minimum cost network flow (FCMCNF) model, can therefore be used to find a practical and complete network solution cost that includes both the “capital” (one-time) costs and the “operating” recurring (variable flow) costs.

Following is the formulation for the capacitated fixed charge minimum cost network flow model. The following mathematical statement defines the optimization problem of finding the minimum-cost set of directed arc flows, with fixed charges, to
route all data through the network:

**FCMCNF:**

\[
\min \sum_{(i,j) \in A} c_{ij} x_{ij} + f_{ij} y_{ij} = z
\]  

\[\text{s.t.}\]

\[
\sum_{(i,k) \in A} x_{ik} - \sum_{(k,j) \in A} x_{kj} = b_k, \forall k \in N
\]  

\[
x_{ij} \leq u_{ij} y_{ij}, \forall (i,j) \in A
\]  

\[
0 \leq y_{ij} \leq 1, \forall (i,j) \in A
\]  

\[
y_{ij} \in \{0,1\}, \forall (i,j) \in A
\]  

where variables are defined as for the MCNF model, plus

\(y_{ij}\) is a binary variable with 1 representing flow is present on an arc \((i,j)\) and 0 indicating the arc has 0 flow, and

\(f_{ij}\) is a one-time fixed-charge scalar for arc \((i,j)\).

Each \(y_{ij}\) is a integer (binary) variable that when set to 1 indicates that traffic can flow on the arc, 0 otherwise. By the use of this modeling technique, arcs with flow can have an associated fixed charge \(f_{ij}\) added to the overall cost. The inclusion of this binary constraint, however, requires solution techniques known as mixed integer programming [36]. The inclusion of these constraints, in (2.12), transforms a linear program into a mixed-integer linear program (MIP), which can be challenging to solve.
2.3.4. Conditional Lower Bounds and Interval Flow Networks

Capacitated, fixed-charge network flow models are useful in aiding practitioners in finding solutions to network flows and costs for many practical problems. One other addition that allows practitioners to improve solutions is the incorporation of a conditional lower bound on arc traffic, thereby constraining arcs to a minimum level of flow or zero flow. The flow on the arc is therefore constrained to be zero or within an interval between some lower and some upper bound. This application of such semi-continuous variables to networks is named “interval flow networks.” It was originally proposed by Barr [5], who, with other researchers, developed heuristics capable of solving large-scale instances [9, 16, 27].

The following is the formulation of the Interval Flow Network Model [5]:

**IFNM:**

\[
\begin{align*}
\text{min} & \sum_{(i,j) \in A} c_{ij}x_{ij} = z \\
\text{such that:} & \\
\sum_{(i,k) \in A} x_{ik} - \sum_{(k,j) \in A} x_{kj} = b_k, \forall k \in N \tag{2.14} \\
y_{ij} \ell_{ij} \leq x_{ij} \leq u_{ij}y_{ij}, \forall (i,j) \in A \tag{2.15} \\
y_{ij} \in [0,1], \forall (i,j) \in A \tag{2.16}
\end{align*}
\]

where \( \ell_{ij} \) is the conditional lower bound on the traffic flow on arc \((i,j)\) in arc set \(A\). The addition of the conditional lower bound adds to the model the valuable feature of putting a constraint on, not only the upper bound of flow on an arc, which is a common equipment capacity limit, but also on the lower bound, which adds the ability to set a minimum arc utilization. This allows for the reduction in the number of active arcs in equipment or links with *de minimis* traffic. Stated differently, this allows designers to “dial-in” the minimum amount of traffic desired on active arcs.
This added capability allows for more advanced and practical SGCN models and solutions; however, it comes at the expense of a more challenging problem to solve.

2.3.5. Fixed-charge Interval-Flow Network Model

Combining the above techniques, a model is now developed that simultaneously addresses practical smart-grid network problems that require capacity constraints, fixed-charges, and lower bounds. The following is the formulation of the Fixed Charge Interval Flow Network Model.

**FCIFNM**:

\[
\min \sum_{(i,j) \in A} c_{ij}x_{ij} + f_{ij}y_{ij} = z \tag{2.17}
\]

s.t.

\[
\sum_{(i,k) \in A} x_{ik} - \sum_{(k,j) \in A} x_{kj} = b_k, \forall k \in N \tag{2.18}
\]

\[
y_{ij} \ell_{ij} \leq x_{ij} \leq u_{ij} y_{ij}, \forall (i,j) \in A \tag{2.19}
\]

\[
0 \leq y_{ij} \leq 1, \forall (i,j) \in A \tag{2.20}
\]

\[
y_{ij} \text{ integer, } \forall (i,j) \in A \tag{2.21}
\]

With this final model, now the formulation of a complete smart-grid communications network model is developed.

2.3.6. Formulating the Complete SGCN Model

Given the FCIFNM formulation described above, it is instructive at this point to view an example model of the SGCN problem in graphical form, as shown in Figure 2.4. When graphing a SGCN, it is typical to draw layers of a hierarchical communications network, to highlight different types of infrastructure and locations involved in the layout of a system. For this example, the first “layer” in the graph is the left-most column of nodes representing homes with smart-meters that generate data to be transported by the network. Smart-meter data traffic from each home
is depicted as inputs to the graph, $b_1$ to $b_8$. The second column of nodes represent equipment called *repeaters* that receive the signals from the smart-meters, amplify and relay those signals further “upstream” in the network. (In practice such repeaters are typically placed on utility poles and other available utility-controlled structures.)

The third row of nodes represent the second type of SGCN equipment called *collectors*, which are aggregation points that concentrate smart-meter traffic for more efficient transport to a destination, such as a data center. Collectors are typically located in a more secure location of the utility, such as fenced-in electric substations. Finally, the fourth column in the graph is the destination of all smart-meter traffic, the data center(s). There the sum of all traffic is taken out of the graph, as represented by the demand at that node.

This example is limited to a small 15-site network, with eight homes, four repeaters, two collectors and one data center. In practical examples, not all locations
can communicate to one another. In this small example all sites are show fully interconnected (though not all arcs are shown for clarity). The eight “homes” supply $b_n$ flow of smart-meter traffic each into the network. The four repeaters collect traffic from the homes closest to the repeater, typically, and deliver that traffic upstream to collectors. In some cases, homes may connect to a more distant repeater, should there be non-line-of-sight (NLOS) between the home and the closest repeater. Similarly repeaters connect to collectors with the strongest signal, not necessarily the closest collector. Also, homes that happen to be close to a collector may bypass the repeater entirely. Lastly the collectors connect to the data center, either directly or via a “backhaul” capability such as the cellular network or a private point to point network.

This is a simplified graph of an SGCN problem in that only the upstream traffic is modeled. Also, with so few smart-meters, there is no need for additional higher capacity or longer-range equipment. That would be needed in other cases where one would need to haul traffic from many disparate collectors back to one or more the data centers far away. This additional equipment can form what is known as the backhaul layer, which can be either cellular service providers, microwave or optical fiber equipment, and is not shown here.

**Constrained Nodes**

While the above forms the basis of a model to characterize a SGCN, another useful technique is needed to further advance modeling real networks: the need to allow for equipment with capacity constraints. In classical models, arcs can be assigned flow values, flow variable costs, arc-establishment fixed costs, and flow capacities in terms of upper and lower bounds. Nodes, on the other hand, do not have such attributes and are simply points of connectivity for arcs. Hence, single nodes that are meant to represent equipment at a physical site cannot be attributed costs, capacity constraints
nor any of the attributes useful to designers who must model real physical equipment at a physical site.

A modeling technique useful for allowing designers to allocate costs and capacities to nodes is the notion of a “constrained node” or more typically called a “split-node”. This technique is useful for modeling practical networks with capacity-limited equipment at various physical sites. This can be achieved in a model using the following technique.

Figure 2.5 depicts a split-node. This name implies that a single node, the large oval shown, is represented by two nodes, the two circles inside, and a connecting arc. The internal arc allows for all the attributes of arcs discussed so far, namely the levying of costs and capacity constraints, to the split node structure itself. As shown in the figure, all arcs into the original node now connect to node $i$ and all arcs out of the original node leave from node $j$. By drawing a larger oval about these two nodes in the model, one can convey the sense of a “constrained node,” representing the equipment at a site. In this way, a limit of flow can be placed on this site node’s equipment,
thereby modeling equipment that has costs and capacity constraints. This cost can even include a net-present-value lump-sum cost of ongoing maintenance summed over a planning horizon of, for example, ten years.

Closest Neighbors

Another concept important to smart-grid design that must be implemented in solving realistically-sized networks is the concept of “closest neighbors.” In practical networks, a particular site’s equipment can only connect to another site’s equipment if the two sites are within range of the technology being used. Hence a model need only include arcs between nodes that are reachable in practice, possibly limited to a set number of the closest nodes. This is done by pre-processing the set of all arcs that could be put into the model, sorting on the physical distance between arcs’ endpoint nodes, then using only a limited set of arcs to the closest neighbors.

The way to depict on a model graph a model that limits the number of arcs emanating from a node is shown in Figures 2.6 and 2.7. For clarity, in this document, instead of showing all arcs emanating from a node, as shown Figure 2.6, a single arc with a special symbol is shown. An arc representing multiple arcs to some closest
neighbor nodes has a *circulation symbol* and parameter indicating the number of closest neighbors, here given as $h$. By limiting the number of arcs from some nodes to some number $h < |N|$, the number of arcs from those nodes grow the total network arcs linearly in $h$, versus geometrically in $h^2 - h$, as would be the case for a fully connected network.

*Upstream only*

It should be mentioned why the approach taken here is to model only the upstream traffic from meters to the data center. Note that adding downstream traffic from the data center to meters would be the case in real SGCNs. In SGCNs the vast majority of the traffic, during typical, nominal, steady-state operating conditions, is upstream. Typical upstream traffic consists of a large volume of data associated with delivering electric power (such as voltage, current, power, power factor, power quality, consumption, and demand). The downstream traffic is typically much less voluminous, consisting of operational, maintenance or infrequent traffic (meter-read and disconnect commands, for example). Only occasionally can some downstream traffic types be significant (such as software updates and upgrades). Thus, the time-averaged, nominal, operating upstream flow of traffic is much greater than downstream traffic. One can approximate downstream traffic to be much less than upstream traffic. Thus, downstream flows, in many cases, can be considered negligible relative to upstream. If however the downstream traffic is significant, then another modeling technique can be used.

Recall the units of traffic flow are “meter’s-worth of traffic.” Typically, a single meter only communicates intermittently in “bursts” of upstream traffic (traffic in the direction of the data center). Likewise, a data center communicates downstream to meters in smaller messages such as acknowledgements, and less frequently in bursts or files (i.e. software upgrades). The upstream traffic is typically much more than the
downstream traffic. Thus one can interpret the upstream one meter’s-worth-of-traffic as a time-average upstream portion and average downstream portion, that flows, on average along the exact reverse path. While this assumption is clearly not true in packet networks (such as IP networks) for every packet in all streams, it again may be a sufficiently acceptable approximation for a basic network dimensioning effort, where a first-order approximate sizing of the network is sufficient. For example, for the one unit of traffic (100%) input into the network by one smart-meter, this can be interpreted to mean, for example 0.7 (70%) upstream and 0.3 (30%) downstream. Then in the solution of the network, the graph of arc flows are interpreted to be 0.7 (70%) upstream traffic, with a mental superposition of an exactly equivalent downstream graph comprised of 0.3 (30%) downstream traffic. This essentially converts the resultant graph from directed arcs into bidirectional links. As mentioned, this assumes downstream traffic flows in the exact same same path as it does upstream. Should this not be acceptable, modeling a bi-directional network is easily extended from the work done here on modeling an upstream-only network.

The above model formulations, each of increasing practicality and complexity, come to form the basis for modeling SGCNs. The classical transportation and transshipment models form the basis of modeling basic communications networks. Adding capacity constraints aids practitioners in modeling practical SGCN by being able to model capacity-limited equipment. The technique of adding fixed charges to arcs further supports designers in modeling practical networks. The use of a lower bound adds a useful technique for eliminating de minimis arcs. The addition of a few other modeling techniques and pre-processing steps add all the tools needed to model practical SGCNs. The next section uses these techniques to formulate a realistic SGCN Network Model.
2.4. The SGCN Model

The formulation of a complete SGCN model requires the use of the above classical forms, with all the aforementioned techniques, (cost minimization, fixed charges, constraints on arcs and through nodes, equipment capacity upper and lower bounds.

Refer to the graph of an SGCN model, shown in Figure 2.8.

Figure 2.8: SGCN Model Graph

2.8. The four major node columns indicate the four types of locations or sites that could be used for SGCN equipment. Homes with smart meters are to be connected back to the data center through a network of equipment placed on utility poles and electric substations. Homes and data centers are typically given in a design problem. The objective of the problem is to determine which of the candidate locations (poles and substations) to use in the final network design.
In this model, meters, repeaters, and collectors are depicted as split-nodes, representing capacity-restricted equipment at this location. These nodes have an internal arc connecting representing capacity-restricted equipment at a location. These nodes have an external arc connecting the device (meter, repeater, or collector) at one location to other devices at different locations. There are two types of external arcs. The first type of external arc connects one type of device to a similar type device, for example a meter to a neighboring meter (or a repeater to a neighboring repeater, etc.). The second type of external arc connects one type of device to a dissimilar-type device, for example a meter (repeater) to a repeater (collector).

In practice, one meter can deliver its own traffic, and it can relay traffic from neighboring meters. The relaying of data through a device is termed meshing. Thus a meshed network of meters means meters can be fully interconnected, in practice, the meshes form around closest neighbors and do not connect network-wide. In general, collector traffic is non-meshed and connected back to the data center via direct point-to-point links.

Setting the number of closest neighbors, whether devices are meshed or non-meshed, can be a design decision based on knowledge of the range of a particular device. The number of closest meter-to-meter neighbors in Figure 2.8 is denoted by parameter $k_1$. The parameter $k_2$ represents the number of closest meter-to-repeater neighbors. Similarly $k_3$ and $k_4$ are parameters for repeater-to-repeater and repeater-to-collector closest neighbors. Collectors all connect back to the data center.

In Figure 2.8, only four homes with meters are shown, but the formulation can be extended to hundreds, thousands, or more homes. Similarly, candidate poles may be all utility poles in the jurisdiction. And all substations can range into the hundreds for large markets.

Sets of nodes and arcs are now defined in order to formulate the mathematical model for the SGCN model, shown in Figure 2.9.
Figure 2.9: SGCN Model Graph with Sets

Node set definitions:

$M_1$ = ordered set of meter-inbound nodes for receiving originating data plus arc flows from other meters’ nodes in set $M_2$

$M_2$ = correspondingly ordered set of meter-outbound nodes for arcs transmitting flow to repeater nodes and to other meters in set $M_1$

$R_1$ = ordered set of repeater nodes for inbound arcs from node sets $M_2$ and $R_2$

$R_2$ = correspondingly ordered set of repeater nodes for outbound arcs transmitting flow to collector nodes in set $C_1$ and to other repeaters in set $R_1$

$C_1$ = ordered set of collector-inbound nodes for receiving arc flows from node sets $R_2$ and $C_2$

$C_2$ = corresponding ordered set of collector-outbound nodes for arcs transmitting flow to the DC node

$D$ = set containing one data-center node

$N = M_1 \cup M_2 \cup R_1 \cup R_2 \cup C_1 \cup C_2 \cup D$


**Arc set definitions**

The model assumes a single data center although extension to multiple data centers is straightforward. In these descriptions, a neighborhood, \( N(i, S, k) \), is the set of up to \( k \) nodes from some node set \( S \) that are closest to node \( i \), where closeness is defined by the geographical distance over which a communications connection can be reliably maintained.

\[
\begin{align*}
A_1 &= \{ (i, j) \mid i \in M_1, j \in M_2, \text{ord}(i) = \text{ord}(j) \} \\
A_2 &= \{ (i, j) \mid i \in M_2, j \in N(j, M_1, k_1), \text{ord}(i) \neq \text{ord}(j) \} \\
A_3 &= \{ (i, j) \mid i \in M_2, j \in R_1, j \in N(j, R_1, k_2) \} \\
A_4 &= \{ (i, j) \mid i \in R_1, j \in R_2, \text{ord}(i) = \text{ord}(j) \} \\
A_5 &= \{ (i, j) \mid i \in R_2, j \in N(j, R_1, k_3), \text{ord}(i) \neq \text{ord}(j) \} \\
A_6 &= \{ (i, j) \mid i \in R_2, j \in C_1, j \in N(j, R_2, k_4) \} \\
A_7 &= \{ (i, j) \mid i \in C_1, j \in C_2, \text{ord}(i) = \text{ord}(j) \} \\
A_8 &= \{ (i, j) \mid i \in C_1, j \in D \} \\
A &= \bigcup_{i=1}^{8} A_i
\]

where scalars \( k_1, \ldots, k_4 \) define the neighborhood sizes and, hence, the density of the network model. For notational convenience, set \( A_0 = A_1 \cup A_4 \cup A_7 \), the set of arcs connecting “split nodes.”

Given the above sets, the mathematical formulation of the general smart-grid model is as follows.
Mathematical Formulation, Smart Grid Model, SG1

\[
\text{SG1 : min } \sum_{(i,j) \in A} C_{ij} x_{ij} + \sum_{(i,j) \in A_0} F_{ij} y_{ij} = z_1 \tag{2.22}
\]

subject to:

\[
1 + \sum_{(j,i) \in A_2} x_{ji} = \sum_{(i,k) \in A_1} x_{ik}, \forall i \in M_1 \tag{2.23}
\]

\[
\sum_{(j,i) \in A_1} x_{ji} = \sum_{(i,k) \in A_2 \cup A_3} x_{ik}, \forall i \in M_2 \tag{2.24}
\]

\[
\sum_{(j,i) \in A_3 \cup A_5} x_{ji} = \sum_{(i,k) \in A_4} x_{ik}, \forall i \in R_1 \tag{2.25}
\]

\[
\sum_{(j,i) \in A_4} x_{ji} = \sum_{(i,k) \in A_5 \cup A_6} x_{ik}, \forall i \in R_2 \tag{2.26}
\]

\[
\sum_{(j,i) \in A_6} x_{ji} = \sum_{(i,k) \in A_7} x_{ik}, \forall i \in C_1 \tag{2.27}
\]

\[
\sum_{(j,i) \in A_7} x_{ji} = \sum_{(i,k) \in A_8} x_{ik}, \forall i \in C_2 \tag{2.28}
\]

\[
\sum_{(j,i) \in A_8} x_{ji} = b, \forall i \in D \tag{2.29}
\]

\[
\ell_{ij} y_{ij} \leq x_{ij} \leq U_{ij} y_{ij}, \forall (i,j) \in A_0 \tag{2.30}
\]

\[
0 \leq y_{ij} \leq 1, \forall (i,j) \in A_0 \tag{2.31}
\]

\[
\text{integer } y_{ij}, \forall (i,j) \in A_0 \tag{2.32}
\]

where \( b = |M_1|, U_{ij}, \) and \( \ell_{ij} \) are, respectively, the upper bound and the conditional lower bound on the flow \( x_{ij} \) on arc \( (i,j) \in A_0 \), and \( C_{ij} \) and \( F_{ij} \) are the variable and fixed costs associated with flow \( x_{ij} \) on arc \( (i,j) \).

The above is the general model for solving smart-grid communications networks with meters communicating upstream to a NOC through repeaters and collectors. All meters and the NOC will be in the solution, as well as repeaters and collectors needed for a cost-minimized network.

This chapter established a careful definition of the problem, with a general model,
SG1, suitable for solving for a smart-grid network. The SGCN model was formulated from classical network flow models. From there a class of instances were set up, shown here, but based upon rules of experimentation discussed next. The general SG1 model was modified slightly for experimentation on a particular scenario, parameters and constants assigned, and these conditions were made ready for problem instances to be generated and solved. In the next chapter the focus turns to an experimental design, and how the above factors and levels were derived. Experimentation quantifies the extent that input factors of Meters, Area, Equipment Price and Lower Bound all play into performance of the model’s optimized designs as reflected in the responses of total network cost (TNC), equipment count (EC), and average arc distance (ALD), all to be discussed next.
Chapter 3
EXPERIMENTATION

In preceding sections, this report described the problem of designing the lowest-cost network topology that connects meters to a data center. Then a mathematical model of the general problem was formulated.

In this chapter, a designed experiment is described and the general model, SG1, is applied to a typical scenario: a suburban smart-meter deployment. Problem instances are generated, used in experimentation, and evaluated to draw conclusions useful to practitioners.

3.1. Experimental Test Problems

Exact real-world data for an operating network could not be obtained due to the confidentiality, security, and proprietary nature of utility customer-data, infrastructure locations, and vendor equipment information. Therefore, in the computational experiments that follow, test problems were created that closely reflect those encountered by practitioners. The scenario and test problems developed mimic a real-world situation and its characteristics. The scenario is described here at a high level, and further explained more accurately in Section 3.8 after precise terms for the experimental design are defined. The scenario is that of a smart-meter deployment in a suburban environment.

Test problems are created for this scenario using a two-step process. In the first step, a network generator takes a set of input parameters (listed and further described in Section 3.6, Table 3.4), and outputs a large network of candidate nodes and arcs. In the second step a solver takes the large candidate network as an input and solves for an optimal lowest cost network.
The code for the generator is provided by Jones [38], and developed from a variation of NETGEN [43], called AMSnetBuild. The solver is called IFNET and further details can be found in Jones [38].

The problem generator and solver combination takes as input the quantities of equipment (meters, available repeaters, and available collectors), and begins by placing them on a square grid area distributed normally (Gaussian) from the center. It also randomly places a data center on the grid to collect the traffic generated by the smart-meters. The area of the grid has a length and width measured in unitless dimensions of “grid points.” The equipment is placed with a higher concentration at the center of the area and with lower density towards the boundaries of the area. This models the non-uniformity of devices across a typical zone. The generator then takes inputs for the network parameters: repeater and collector equipment prices, variable flow costs, and upper and lower bounds on arcs.

The complete network problem, consisting of quantities of equipment, capacity constraints, and prices, are then input to the IFNET solver. The solver code computes the optimal (lowest cost) network connectivity from the given the meters and the available repeaters and available collectors, used for traffic aggregation point, and prices of equipment.

Definitions of experimental terms are provided next. This includes input factors and output response variables.

3.2. Response Variables and Performance Evaluation Criteria

The design of a network typically starts with an area to cover and services to be provided. The design uses network-infrastructure equipment selected by, or provided to, the practitioner. Ultimately, a network design results in network characteristics that determine the “goodness” of the design.
These response variables include system evaluation attributes such as the cost of the network, the amount of equipment used, and the nature of the communications links (i.e., the average distance of the links). What follows are the important responses for the present study.

3.2.1. Total Network Cost

The total cost of a network typically includes the sum of the costs of equipment, but can also include site costs such as real estate and infrastructure (cabinets, antenna structures, and ancillary material) and even the net present value of recurring support and maintenance costs. Usually, the cost of the smart-meter itself is accounted for in other budget allocations, and is not included in the costs of the communications network itself. This allows separation of the meter purchase decision from the communications network purchase decision. For purposes of this study, the Total Network Cost (TNC) is defined as the cost of the network equipment and material but not the cost of the meters, meter radios, data center equipment, site real estate, or support and maintenance of the network. Thus, Total Network Cost includes the cost of the network equipment, specifically repeaters and collectors and their ancillary material needed for operation.

Another “cost” included in the Total Network Cost is a value reflecting the length of a communications arc. This cost is not an amount in real dollars, rather a “virtual” or “distance” cost that has been added to the Total Network Cost to account for latency or reliability of the connection, since forcing shorter arcs over longer arcs in a solution is useful because the longer the arc distance for wireless equipment, the lower the signal strength and the lower the link reliability. By minimizing both real-cost and distance-cost, the result is a network design that is both cost-efficient as well as more reliable. In this experiment, distance cost is included in the Total Network Cost by way of the variable flow cost for arcs between meters, repeaters and collectors, and is modeled as one dollar per unit-distance.
3.2.2. Equipment Count

The network considered here is the set of repeaters and collectors that feasibly connect the given number of smart-meters to a data center. Equipment Count (EC) is the total number of repeaters and collectors in a specific design. Equipment Count for each design instance is characterized by enumerating the number of repeaters and collectors used.

This is an important characteristic of a network design since the less equipment required for the network solutions, typically the lower the overall cost of the network and ease of maintenance. In practical problems, the increased equipment count drives up cost for both the equipment itself and the associated site real-estate leases, and equipment maintenance. Reducing Equipment Count is one way of reducing related indirect costs not accounted for in the total network cost metric. For example, indirect costs are associated with deploying, operating and maintaining a larger network. It is useful to minimize the quantity of ancillary materiel and resources at sites such as antennas, equipment cabinets, and electrical power needed. Additionally this fosters “green network” initiatives that seek to minimize the amount of sites with unsightly “urban furniture” placed in the environment. By knowing what relates to minimizing Equipment Count, practitioners can support green-planet designs. These experiments collect Equipment Count as an important response variable.

3.2.3. Average Link Distance

Average link distance (ALD) provides a view into an important characteristic of a particular network design. As previously mentioned, longer average links translate into a slight or sometimes significant issue with link reliability. The longer the links in a network design the lower the signal level available for detection, and the more likely the link can be perturbed by conditions in the environment.
Described above are the important outcome characteristics of a network design: Total Network Cost, Equipment Count and Average Link Distance. A “good” design is an attractive balance of all three. How to achieve a good set of output responses is a function of the inputs (namely, the problem situation given) and the design decisions made. By experimentation with a large number of combinations of input factors, the experimenter can become informed of the output Total Network Cost, Equipment Count and Average Link Distance responses, and search for the combinations of factors that lead to attractive responses. These inputs are described next.

3.3. Factors and Levels to be Explored

In the design of a network, the designer typically has some problem “givens,” such as the area to cover, but then is free to use experience and judgment to specify the type of equipment and the placement of the equipment to form a communications network. The area to cover typically sets the number of meters within the area, and those two criteria are typically pre-determined. The decisions made concerning the location and type of equipment are design decisions. In selecting the equipment, the practitioner must specify operating characteristics, including minimum and maximum traffic levels and capacity capabilities of the network links. The maximum setting is a constraint set by the equipment’s capabilities. The minimum constraint can be a design decision by the practitioner to not place equipment when the flow through it is below some minimal level. This amounts to discouraging arc placement from a site for a small amount of flow when that flow might be diverted elsewhere, thus saving on Equipment Count and eliminating equipment of low-traffic-utilization arcs.

*Meters* are the number of electric smart-meters in the experiment. Equivalently it is the number of households in the service area that require a smart-meter. The levels shall be set to address a small town or city zone deployment. These levels are set to 5000 and 10,000 *Meters.*
**Area** is the second factor, and it is the range of service coverage for a particular deployment. *Area* levels are measured in unitless grid points. The user inputs a unitless length and width, for example, a width of 100 and length of 100. Then a network generator takes the number of Meters and places them within the area using a Gaussian density function to represent a higher density of Meters at the center of the area. The *Area*, when scaled and taken with the levels of *Meters* above, provides a density of meters per square area. The *Area* quantities selected for experimentation are (100x100) and 500x500. These grid-point Areas can be scaled from unitless gird point to a distance unit such as miles, to provide a density representative of a suburban area.

**Lower-Bound** is the third factor. It is the Lower-Bound, of minimum traffic flow for an arc to be used or activated. This is useful when decisions must be made to use or not use equipment if the traffic though the equipment is lower than some threshold. This is done by placing these Lower-Bound arcs on the equipment split nodes, thereby only allowing equipment to be part of the solution that are above the Lower-Bound. By setting the Lower-Bound to a high level, more traffic is diverted to fewer arcs with more flow, resulting in fewer arcs overall, and possibly resulting in fewer sites. For experimentation, the levels selected are 20 and 200 units of flow. One unit of flow is defined as “one meter’s worth” of data traffic.

**Equipment-Price** is the fourth factor. Clearly the cost of the equipment, specifically the cost of the repeaters and collectors, has a direct impact on the Total Network Cost. However, it is of interest to see how this factor has an effect on the response when combined with the other factors. Furthermore it is of interest to determine how Equipment Price (EP), residing in the fixed charge of internal equipment arcs, ranks in terms of influencing the response, versus how the “virtual cost,” residing in the external arcs’ variable cost, ranks in relationship to the response. For experimentation, the levels selected are an order of magnitude apart: $20,000 and $200,000. These levels include the devices themselves plus costs to deploy.
3.4. The Hypothesis to Be Tested

In statistical experimentation, an experimenter postulates one or more hypotheses about how a system responds to various inputs, and wishes to confirm or reject those conjectures. Or an experimenter may wish to create a statistical model of a system by statistically characterizing how the system responds to a certain set of inputs, thus developing a statistical model useful for prediction. The experiment is then designed by specifying the appropriate conditions of the experiment and the set of inputs, termed factors, and the set of outputs, termed responses. In these computer experiments, a solver is the term used to describe the software that computes the results of a particular set of inputs to the experiment, using the mathematical model earlier established. Thus, an experiment consist of setting all factors at specific levels, running the solver, and noting the responses.

In general, there are several conjectures that can be made concerning how factors are related to responses in the present study. For example, obvious conjectures include: “The Equipment-Price, number of Meters, and Area are all potentially related to the total network cost.” More subtle conjectures include “The number of Meters likely has an effect on the Equipment Count,” and “The Lower-Bound likely has an effect on Average Link Distance.” In order to confirm or reject these conjectures, each conjecture is subjected to a hypothesis test. There are 30 such conjectures that lead to 30 hypotheses to be tested in this research. In this research 30 hypotheses are tested shown in Table 3.3. Hypothesis testing is covered in detail in several classic texts on design of experiments and statistics [52, p. 36; 63, p. 4-1; 70, p. 284].

The following terminology is used in order to describe the hypothesis tests in terms of the execution of the solver software. An experiment is typically run by setting one factor at one level and observing one output response. Thus one observation is made from this one experimental run. In this work, 30 hypotheses are posited, and this typically would require 30 experimental runs and 30 observations.
Table 3.3: Hypotheses to be Tested

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ((H_0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The factor Meters has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>2</td>
<td>The factor Area has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>3</td>
<td>The factor Lower-Bound has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>4</td>
<td>The factor Equipment-Price has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>5</td>
<td>The factor Meters*Area has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>6</td>
<td>The factor Area*Lower-Bound has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>7</td>
<td>The factor Meters*Equipment-Price has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>8</td>
<td>The factor Area*Lower-Bound has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>9</td>
<td>The factor Area*Equipment-Price has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>10</td>
<td>The factor Lower-Bound*Equipment-Price has no effect on the response Total Network Cost</td>
</tr>
<tr>
<td>11</td>
<td>The factor Meters has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>12</td>
<td>The factor Area has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>13</td>
<td>The factor Lower-Bound has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>14</td>
<td>The factor Meters*Area has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>15</td>
<td>The factor Area*Lower-Bound has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>16</td>
<td>The factor Meters*Equipment-Price has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>17</td>
<td>The factor Area*Lower-Bound has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>18</td>
<td>The factor Area*Equipment-Price has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>19</td>
<td>The factor Lower-Bound*Equipment-Price has no effect on the response Equipment Counts</td>
</tr>
<tr>
<td>20</td>
<td>The factor Meters has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>21</td>
<td>The factor Area has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>22</td>
<td>The factor Lower-Bound has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>23</td>
<td>The factor Meters*Area has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>24</td>
<td>The factor Area*Lower-Bound has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>25</td>
<td>The factor Meters*Equipment-Price has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>26</td>
<td>The factor Area*Lower-Bound has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>27</td>
<td>The factor Meters*Equipment-Price has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>28</td>
<td>The factor Area*Lower-Bound has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>29</td>
<td>The factor Area*Equipment-Price has no effect on the response Average Link Distance</td>
</tr>
<tr>
<td>30</td>
<td>The factor Lower-Bound*Equipment-Price has no effect on the response Average Link Distance</td>
</tr>
</tbody>
</table>
In this work however, the solver software is designed such that four factors are required as inputs to compute a solution, and each solution is characterized by three responses. Thus, one set of four input factors results in one solution comprised of three output responses. In this work, 16 instances (executions) of the solver are performed, resulting in 16 three-response results, or 48 responses total.

The hypotheses to be tested include main effects and interactions. In Table 3.3, the first four hypotheses (numbers 1-4 in the table) are main effects, and the next six hypotheses (numbers 5-10) are the interactions of the four main effects. Tests of interactions, shown in the form Factor1*Factor2, and read “Factor 1 crossed with Factor 2,” are conducted when it is believed that the response has an effect on not only by one factor’s level, but also in combination with a second factor, and its level. The interaction hypotheses (numbers 5-10) are therefore Meters*Area, Area*Lower-Bound, Meters*Equipment-Price, Area*Lower-Bound, Area*Equipment-Price and Lower-Bound*Equipment-Price.

3.5. The Design

Given the input factors, output responses, and hypotheses to be tested, what remains is the careful design of the experiment that will collect the appropriate set of data for statistical analysis. Setting the number and level of each of the input factors for the experiment is important. Proper selections provide advantages during analysis and improper selections may provide inaccurate or incorrect conclusions. An experimental approach termed “best guess,” whereby various levels are set for various factors based on the experimenter’s knowledge, interest or intuition, is only useful should the experimenter have significant knowledge of the behavior the factors have on the responses [52, p. 4]. But this approach may still lead to errors. One-factor-at-a-time selection of factor values is another approach that sets a baseline of factors, and tests each factor separately holding all other factors constant.
This approach however does not take into account interactions. The preferred approach, conducted here, is to perform a *full-factorial* set of experiments. A full-factorial set of experimental data allows for the following important advantages [52, p. 186]: it avoids misleading conclusions when interactions exist between factors in the data; full-factorial data sets are more efficient in terms of experimental runs needed, versus one-at-a-time experiments; and the effects of a factor can be estimated at several levels of the other factors. The fact that the factorial data set is full, with no missing data, ensures that the effects of each factor can be estimated independently of the other factors [63, p.1-44]. The experiment conducted here is a full-factorial experiment, with a full-factorial set of observations.

How does the experimenter decide if a factor has an effect on a response? More critically, if there is an effect to some degree, is there a standard threshold an experimenter would use to deem an effect “significant.” In statistics this significance level is referred to as $\alpha$ and is determined prior to data collection and analyses. To select this alpha value one must understand the concepts of comparing sample means and the errors one wishes to avoid in hypothesis testing. Comparing sample means is simply taking the mean of the response of an experiment that include observations with the factor in question and comparing it with the mean of the response when the observations that do not include the factor in question. A statistical means-comparison test could be a *t*-test or other *test statistic* (e.g., $F$-test), discussed below. In hypothesis testing, if an experimenter judges a hypothesis to be true, and it is indeed true, there is no error. But if an experimenter judges a hypothesis to be false and it is actually true, the experimenter has committed a “Type I” error. A Type I error is when the experimenter has rejected the null hypothesis when it is true. $\alpha$ is the probability of such an error. Alternatively, if the experimenter judges the hypothesis to be true, but it is actually false, the experimenter has committed a “Type II” error. A Type II error is when an experimenter does not reject a false null hypothesis. Such an error is given the special symbol $\beta$. Thus,
\[ \alpha = P(\text{Type I error}) = P(\text{reject } H_0|H_0 \text{ is true}) \]

\[ \beta = P(\text{Type II error}) = P(\text{fail to reject } H_0|H_0 \text{ is false}) \]

where \( \alpha \) and \( \beta \) are thus figures of merit (FoM) for experiments that are often set at probabilities sufficiently small to provide the experimenter with sufficient confidence that Type I and Type II errors are unlikely. A typical value for \( \alpha \) is 0.05, signifying a 95\% probability there is no error in judgment (no Type I error). A FoM for \( \beta \) could be set at a higher level, if making a Type II error is not as critical as as Type I error. So for an experimenter to decide if a factor has an effect on a response, to a statistically significant level, the experimenter would take a random sample, compute an appropriate test static, compare the response with and without the factor in question, and either reject or fail to reject the null hypothesis, based on whether the probability of a Type I error is less than \( \alpha \). In practice, the focus is typically on setting \( \alpha \), testing the null hypothesis, and if the test statistic is computed to be less than \( \alpha \), reject the null hypothesis in favor of an alternative hypothesis, and expect the judgment to be true with a probability of \( (1 - \alpha) \). This type of hypothesis test is formally written mathematically as follows.

\[ H_0 : \mu_A = \mu_B \quad (3.1) \]

\[ H_A : \mu_A \neq \mu_B \quad (3.2) \]

By performing this test on the hypotheses of Table 3.3, conclusions can be drawn as to whether a factor has a statistically significant effect on a response. This is how the hypotheses of Table 3.3 shall be tested.

The statistical analysis used to reject or fail to reject the null hypothesis is analysis of variance (ANOVA) testing. In ANOVA testing, the mean of the response across all observations when the factor level is set at the higher of two levels, is compared to the mean of the response across all observations when the factor level is set at the lower
of two levels. If there are only two levels to test, a $t$-test is performed using the test statistic that follows the Student’s $t$-distribution. This determines whether the two sets of data, at the two levels of the factor, are significantly different from each other. If there is a statistically significant difference in these means, one can conclude there is an effect from that factor on that response. If there is not a sufficient difference between the means of the response between the two levels of the factor, one fails to reject the hypothesis that “the means are equal,” concluding there is insufficient evidence to support the null hypothesis.

If there are more than two means to test a special test, the Tukey HSD test, can be performed. This test can compare all means at once, and place them into statistically different groups. The test of whether there is a statistically sufficient difference is provided by comparing a calculated “$p$-value” (arising from the ANOVA calculations) to the $\alpha$ threshold. If the $p$-value is smaller than the significance level ($\alpha=0.05$) then one can conclude there is a statistically significant difference between means, that the means are not equal, and the null hypothesis is rejected.

Therefore, the design of this computer experiment is to do a full-factorial modeling experiment with two levels per factor, one replication per sample, and to perform ANOVA means tests to establish correlation significance at the $\alpha=0.05$ level for main effects and their interactions, to the responses of interest.

Beyond hypothesis testing discussed above, which rejects or fails to reject a hypotheses posited in Table 3.3, a regression on the data is performed that helps identify the relative importance among all factors. A statistical model is established that contains the main effects and interaction terms. The simplified form of the model is presented in the equation

$$y_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{1i} x_{2i} + \epsilon_i$$

(3.3)

where:

$y_i$ is the the output response
\( \gamma_0 \) is the \( y \)-axis intercept or grand mean of \( y \)
\( \gamma_1 \) is the coefficient of the first factor, or equivalently the slope or factor effect
\( x_1 \) is the first input factor
\( i \) is the \( i \)th observation
\( \gamma_2 \) is the second factor coefficient giving the slope or factor effect
\( x_2 \) is the second input factor
\( \gamma_3 \) is the interaction term coefficient
\( x_1 x_2 \) is the interaction term
\( \epsilon_i \) is the random error.

This is a second-order model due to the existence of both main terms and interaction (multiplicative) terms. The regression analysis seeks to determine the \( \gamma \) coefficients in the model that best fits the sample of input observations to output responses. A coefficient close to or equal to zero indicates a negligible or zero correlation between an input and the output response. A negative coefficient indicates a negative correlation between an input and the output. Thus the magnitude of the coefficient indicates the effect of the factor input on the response output. The model is a predictive equation that predicts the output, \( y_i \), given a set of inputs \( x_i \). The above model is useful for prediction and, given theorized input settings not observed by experiment, the equation can be used to predict output responses through the use of the statistical model obtained.

3.6. The Experimental Scenario

The scenario used for this experimentation is as follows. A neighborhood Area is taken with an assumed number of Meters, and a set of design decisions established for the experiment regarding the Lower bounds on equipment, and the price of the equipment.

The experiment is designed to take place in a suburban setting. Ranges of the
factors are scaled such that they are representative of various household densities that comprise suburban areas. Density is not a main factor, rather it can be observed as the combination of the factors of *Meters* and *Area*. Table 3.4 shows the calculation of Area in square miles for both *Area* factor levels (100 and 500 grid points square).

Table 3.4: Calculation of Areas in Square Miles

<table>
<thead>
<tr>
<th>Metric</th>
<th>Area Factor Level 1</th>
<th>Area Factor Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Length (Unitless)</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Edge Lengths (Feet)</td>
<td>118.064</td>
<td>118.064</td>
</tr>
<tr>
<td>Area (feet)</td>
<td>11,806.438</td>
<td>59,032.188</td>
</tr>
<tr>
<td>Edge Lengths (Miles)</td>
<td>2.236</td>
<td>11.180</td>
</tr>
<tr>
<td>Area (Sq. Miles)</td>
<td>5.000</td>
<td>125.000</td>
</tr>
</tbody>
</table>

From the above Area in square miles, four levels of densities are derived in Table 3.5. These densities represent a range of suburban Areas from 40 households per square mile to 2000 households per square mile.

Table 3.5: Densities

<table>
<thead>
<tr>
<th>Meters</th>
<th>Area (grid points)</th>
<th>Area (Square Miles)</th>
<th>Density (Meters/Sq. Mi.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>100</td>
<td>5</td>
<td>1000</td>
</tr>
<tr>
<td>5000</td>
<td>500</td>
<td>125</td>
<td>40</td>
</tr>
<tr>
<td>10000</td>
<td>100</td>
<td>5</td>
<td>2000</td>
</tr>
<tr>
<td>10000</td>
<td>500</td>
<td>125</td>
<td>80</td>
</tr>
</tbody>
</table>

As an illustrative example, Figure 3.3 shows a density map of the city of Dal-
las, Texas, USA from the database City-Data.com [20]. Dallas has a wide range of households per square mile. The map shows a range of densities from zero, in remote undeveloped areas, to over 12,000 in the heavy urban center or in areas with large housing developments. Thus, the densities of this study fall into the lower ranges of this map (light blue areas).

Figure 3.3: Household Density, Dallas Texas USA.

As a secondary example, the author’s neighborhood is shown in Figure 3.4. The computed household density is 2,543 homes per square mile (407 homes in an area of 0.16 square miles). This is slightly larger than the densities listed in Table 2 above. Thus, this study addresses neighborhoods that are slightly less dense than that illustrated here.
Overall, modeled here is a network within an area having meters to be connected to a data center. In order to connect all Meters to the data center, the model allows sites (nodes) within the Area to be equipped as either repeaters or collectors. These then form potential aggregation points for upstream traffic, as required, should it lower the overall cost of the solution. These potentially employable candidate sites are termed Available Repeater(s), and Available Collector(s), in the table.

These parameters and factors are collected and listed in Table 3.6. The solver takes as inputs the values shown in the table. The four factors are: Meters (5,000 and 10,000), Area (100, 500), Lower-Bound (20, 200) and Equipment-Price (20,000 and 200,000). The following design parameters are held constant: Available Repeaters, Available Collectors, Variable Costs and Upper bounds.
Table 3.6: Solver Parameter Inputs

<table>
<thead>
<tr>
<th>Factor &amp; Parameters</th>
<th>Constants</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters</td>
<td></td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Available Repeaters</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Available Collectors</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area (grid points on length side)</td>
<td>100</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Area(grid points on width side)</td>
<td>100</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Equipment-Price Repeater</td>
<td>20,000</td>
<td>200,000</td>
<td></td>
</tr>
<tr>
<td>Variable Cost Repeater</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeater Lower Bound</td>
<td>20</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>Repeater Upper Bound</td>
<td>2,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collector Lower Bound</td>
<td>20</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>Collector Upper Bound</td>
<td>16,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equipment-Price Collector</td>
<td>20,000</td>
<td>200,000</td>
<td></td>
</tr>
<tr>
<td>Variable Cost Collector</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.7. The Analysis

The analysis of the experimental data is comprised of an analysis of each of the list of hypothesis tests; tests that are each comprised of two hypothesis, the null ($H_0$) and the alternative hypothesis ($H_1$). The series of tests will reject or fail to reject the null hypothesis. When the null hypothesis $H_0$ is rejected, there is sufficient evidence to support a conclusion that the alternative hypothesis $H_1$ is true, within the significance level of the test. As an example, taking a hypothesis that the level of meters has no effect on Total Network Cost, and rejecting that hypothesis through statistical analysis, one can conclude, within the significance level, that indeed, the level of meters has an effect on (non-causal) the Total Network Cost.

After conclusions are reached on the various hypotheses, a regression analysis is performed on the data. The regression analysis sheds light on the relative strength of the influence of each of the factors and their interactions on the responses. By examining the fitted model equation, one can get a feel for the holistic contribution of influence that all factors have on a particular response.

Experimental tests and analyses were performed on the observations collected. Statistical computations were performed using the JMP [60] application. Specifically, ANOVA testing was performed using JMP for main effects and interactions. Then regression analysis computations were performed using the leased square method in JMP [61].

3.8. The Experimental Data

The set of data captured for this experiment is shown in Table 3.7. It supports a full-factorial experimental design with one observation per factor per level.

While classical experimentation procedures necessitate experimental runs be taken in random order, to minimize the effect of nuisance factors in the environment beyond the control of the experimenter, the response variable values obtained from each observation are identical regardless of the run order, randomization of the run order
was not performed. One observation was taken for each combination of the factor levels, resulting in sixteen total observations.

The computing environment for these runs is a Toshiba Model R94SUS1 PC running Red Hat Enterprise Linux Workstation, release 6.8 (Santiago) with 7.6 GB RAM, a four-core processor (Intel Core i5, 3320 M CPU at 2.6 GHz), and 27 GB of available hard disk. Data is taken by entry of each set of data to the solver, which provided the Total Network Cost value directly. Equipment Count and Average Link Distance were computed from the solution data output.

Table 3.7: Experimental Data

<table>
<thead>
<tr>
<th>Run</th>
<th>Factors</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Meters (X₁)</td>
<td>Area (X₂)</td>
</tr>
<tr>
<td>1</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>10,000</td>
<td>500</td>
</tr>
<tr>
<td>3</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>5,000</td>
<td>500</td>
</tr>
<tr>
<td>5</td>
<td>10,000</td>
<td>500</td>
</tr>
<tr>
<td>6</td>
<td>5,000</td>
<td>500</td>
</tr>
<tr>
<td>7</td>
<td>5,000</td>
<td>500</td>
</tr>
<tr>
<td>8</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>5,000</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>5,000</td>
<td>500</td>
</tr>
<tr>
<td>11</td>
<td>5,000</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>5,000</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>10,000</td>
<td>500</td>
</tr>
<tr>
<td>14</td>
<td>5,000</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>10,000</td>
<td>500</td>
</tr>
</tbody>
</table>
In Table 3.7, a variable is shown for each of the Factors and Responses. The Factor variables take on two levels. The Response variables take on real numbers or integers.

In Table 3.8 Experiment variables are defined for describing the experiments in the following section. They take on two levels for the main effects shown in the table, and four levels for the interaction terms. Table 3.9 lists the Experiment response variables. All these variables appear in Experiments #1 though #30 that follow.

Table 3.8: Experiment Variables - Main Effect and Interaction Terms

<table>
<thead>
<tr>
<th>Experimental Variable</th>
<th>Constituent(s)</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>Meters</td>
<td>2</td>
</tr>
<tr>
<td>$X_2$</td>
<td>Area</td>
<td>2</td>
</tr>
<tr>
<td>$X_3$</td>
<td>Lower-Bound</td>
<td>2</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Equipment-Price</td>
<td>2</td>
</tr>
<tr>
<td>$X_5$</td>
<td>Meters*Area</td>
<td>4</td>
</tr>
<tr>
<td>$X_6$</td>
<td>Meters*Lower-Bound</td>
<td>4</td>
</tr>
<tr>
<td>$X_7$</td>
<td>Meters*Equipment-Price</td>
<td>4</td>
</tr>
<tr>
<td>$X_8$</td>
<td>Area*Lower-Bound</td>
<td>4</td>
</tr>
<tr>
<td>$X_9$</td>
<td>Area*Equipment-Price</td>
<td>4</td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>Lower-Bound*Equipment-Price</td>
<td>4</td>
</tr>
<tr>
<td>Experimental Variable</td>
<td>Response</td>
<td>Type</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>$Y_1$</td>
<td>Total Network Cost</td>
<td>real</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>Equipment Count</td>
<td>integer</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>Average Link Distance</td>
<td>real</td>
</tr>
</tbody>
</table>

3.9. Interpretation of Analysis Computations

The following is an explanation of the statistical-analysis and hypothesis-testing calculations, and the tables and graphs used to analyze the data to make decisions on the hypothesis tested. Each of the following experiments begins with a statement of the hypothesis under test. The form of the hypothesis test is as follows. For an example experiment “A,” both the null hypothesis and the alternative hypothesis are listed for input factor B and output response C. This is the hypothesis under test, such as the following:

Experiment A:

$H_0 \ #A$: The response B is the same for both levels of the factor C.

$H_1 \ #A$: The response B is not the same for both levels of the factor C.

Following the experiment statement, the first item in the analysis is a least-squared means plot (LS Means Plot). Here, as an example, I will refer to a study on pain versus gender. Referring to Figure 3.5, factor levels for the input factor “gender” is shown on the abscissa and the output response “pain” values are shown on the ordinate. For each categorical factor, for example for the factor female, the mean value of all response values is plotted as a point, and brackets about the point indicates the 95% confidence interval range about the values observed for the factor female. If only
one value is available (for experiments with only one replication, such as the present experimentation) the brackets are not shown. This plot is valuable to observe the difference in means and range of the response (pain), for each level (female, male) of the factor (gender). If the line is horizontal, there is no difference between the means (no pain difference), thus there is likely little or no correlation of this factor to the response. Should there be a large difference between the means, resulting in a large slope in the line, this indicates there may be an effect of the factor on the response.

Figure 3.5: LS Means Plot Gender vs. Pain

Next an analysis of variance calculation is performed on the data, and the results are displayed in a table labeled “ANOVA” An example ANOVA table for Meters is shown in Figure 3.10.

Table 3.10: Example ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Mean Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters</td>
<td>1</td>
<td>1</td>
<td>8.031E+11</td>
<td>101</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

This table indicates the probability whether or not there is a difference between
the response means for the two levels of the factor, here Meters. Meters is the Source of variation of the response. The number of replications is listed as number of points \( NP \). Given \( k \) levels of the factor, the degrees of freedom \( DF \) is computed as \( k - 1 \), in this case two levels of the factor Meters gives a \( DF = 1 \). The Mean Sum of Squares, also termed the regression sum of square, or mean square treatment, is the amount of variation in the response values due to the factor (treatment), versus the total variation in the response data, termed the error sum of squares, or mean square error. The \( F \)-Ratio is the ratio of the mean square treatment to the mean square error. This is also termed the \( F \) test statistic. The table value labeled \( Prob > F \) is the computed \( p \)-value. This value is compared to the level of statistical significance, the \( \alpha \) level, to determine if the means are significantly different. In the example experiment A above, if the \( p \)-value is greater than \( \alpha \) we accept the null hypothesis \( H_0 \), otherwise if the \( p \)-value is less than \( \alpha \) we reject the null hypothesis. Using the \( F \) test statistic compared to the \( \alpha \) threshold allows a determination of acceptance or rejection of the null hypothesis.

The \( F \) statistic used above is often used for comparisons of variance between two or more samples. The \( t \) statistic used next is often used to estimate the mean of a population. A two-sided \( t \)-test is used to compare estimated means of two populations. For example, given an experiment that requires comparing two samples, one with data taken with the level of a factor, Meters, is set to 10,000, resulting in a mean of the response variable, in this example Total Network Cost of 1,634,354, and a second set of data with Meters set to 5,000, resulting in a mean of 1,186,274. The result of a two-sided \( t \)-test is shown in Table 3.11. The value \( Difference \) is the difference in the means (1,634,354 - 1,186,274 = 448,081). This value is used in the numerator of the \( t \)-Ratio. The denominator of the \( t \)-Ratio is the standard error, \( Std \ Err \ Dif \) (44,606). Note this is the red marker shown on the plot to the right of the difference of means density curve, to be explained shortly. From these two values the \( t \)-ratio in the table is computed to be 10.04535. The degrees of freedom used in the test is shown as
DF. The $p$-value associated with a two-tailed test (0.0002*) is labeled $Prob > |t|$. An asterisk means that the $p$-value is less than the $\alpha$ value of the test, which would reject the null hypothesis (the means come from the same population). The $p$-value associated with an upper-tailed test is denoted by $Prob > t$. The $p$-value associated with a lower-tailed test is denoted by $Prob < t$.

The $t$-test plot shows the sampling distribution of the difference in the means. The vertical red line on the $x$-axis is the difference in the means $(1,634,354 - 1,186,274 = 448,081)$. The curve is the probability density function (PDF) of the difference of the two sample means. This PDF is formed by subtracting the two sample means to form one new difference-mean and combining the two sample variances to form one new variance [8]. The resulting PDF has probability values on the $y$-axis from 0 to 0.5 and a sample-mean-difference-values along the $x$-axis, termed difference-values hereafter.

Given a desire to identify whether two sets of data are significantly different in their means, one would compute the $t$-test plot described above, (a probability versus difference-values plot), mark a value of statistical significance the difference-values $x$-axis, label it as the critical value, and observe if the difference-value one wishes to compare, falls to left or right of the critical value. In this case, the critical value is set to $\alpha = 0.05$ and statistical significance then would be represented by a $p$-value $\leq 0.05$. Here, the computed difference-value is 448,081, which results in a $p$-value $= 0.0002^*$, leading to a decision that the difference in means is significantly different. This difference-of-two-sample-means plot is useful for observing the degree to which two sets of data are significantly different, by noting how far the red marker is from the center of the plot. The $p$-value is equivalent to the area under the curve beyond the red marker (not easily seen in this plot with such a low $p$-value).\(^1\)

\(^1\)See Table 3.23 on page 91 for an example with a $p$-value much greater than the test's $\alpha$ value, showing a large shaded area, indicating not enough evidence to reject the null hypothesis that the means are equal, thus negating the inference that the factor-levels make a statistically-significant
Table 3.11: Student’s t-test for Meters Comparing 10,000 with 5,000

<table>
<thead>
<tr>
<th>Difference</th>
<th>448,081</th>
<th>t Ratio</th>
<th>10.04535</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>44,606</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>562,743</td>
<td>Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>333,418</td>
<td>Prob &gt; t</td>
<td>&lt; .0001*</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>Prob &lt; t</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

To further visualize the relationship between the factors, a *Letters Report* is created, and an example is shown in Table 3.12. A letters report is a table that assigns a particular factor’s level to a group based on the sample variance of the data for that level. In the table, the set of observations taken for the factor Meters at level 10,000 has a variance statistically different from the variance of the set of observations taken for the factor Meters at level 5,000. Thus, the level 10,000 is placed in a different group, group A, than the level 5,000. The mean of the set of observations taken per level for the response “Total Network Cost” is also shown. Here the mean for level 10,000 is a Total Network Cost of 1,634,354 and for level 5,000 is 1,186,274. From the letters report, on can observe and conclude that the level of meters has a statistically significant effect on the Total Network Cost, since level 10,000 falls into a separate group, than the group for level 5,000.

For a two-level test, a *t*-test is performed. For more than two levels, a Tukey HSD test is used. In the latter case, if one or more factor’s level’s variances are statistically similar, that group is assigned to the same group letter. If one group of factor-level’s variances are statistically different from another group, that factor-level’s group is assigned to a different group letter. For an example of this, refer to Table 3.35 on page 106. In that table, four levels of the interaction Area*Equipment-Price are listed on four rows. Each level has a different Total Network Cost mean difference on the response.
and, since all four levels have a statistically significant variance, they are grouped into four distinct groups A, B, C, and D. Therefore one can visually note that each level is statistically significantly different from each other. The Tukey HSD method allows a quick visualization of the relative differences of groups of observations, and their relative statistically significant differences.

Table 3.12: Letters Report Meters

<table>
<thead>
<tr>
<th>Meters</th>
<th>$t$-test Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>A</td>
<td>1,634,354</td>
</tr>
<tr>
<td>5000</td>
<td>B</td>
<td>1,186,274</td>
</tr>
</tbody>
</table>

Following the hypothesis testing, regression analysis is performed and a regression equation is computed. The regression performed is a *least-squares fit* of the four *main effects* and six *interaction terms* computed from the data. This information is then used to derive a prediction equation.

An example of the regression analysis tabulation is shown in Table 3.13. The table lists the independent term $x$ or variable interaction in the column labeled *Term*. In this example, the first $x$ term, in the first row is Equipment-Price. This is followed by the *Estimate* of the $\beta_i$ coefficient (e.g., 8.6625632), the *Standard Error*, the computed *t-Ratio*, and the *p*-value (*"Prob > |t|"*). The rows of the table are sorted by $t$-ratio value, which highlights the variables that have the most effect on the response.
Table 3.13: Sorted Parameter Estimates

| Term                      | Estimate $\beta_i$ | Std Error | $t$-Ratio | $Prob > |t|$ |
|---------------------------|--------------------|-----------|-----------|---------|
| Equipment-Price           | 8.6625632          | 0.24781   | 34.96     | < .0001* |
| Area                      | 2187.2778          | 111.5145  | 19.61     | < .0001* |
| Meters                    | 89.616125          | 8.921158  | 10.05     | 0.0002*  |
| (Area-300)\*(Equipment-Price-110000) | 0.0097354   | 0.001239  | 7.86      | 0.0005*  |

Next a standard Analysis of Variance table, such as Table 3.14, is presented. In this example, a good fit of the regression line to the data is apparent given the $p$-value ($Prob > F$) of $< 0.0001$* (the asterisk denoting statistical significance), which is the likelihood that the fitted model is significant.

Finally, a Summary of Fit table is shown in Table 3.15. An example set of characteristics is shown with values for this example. The the row labeled “RSquare” shows the coefficient of determination, denoted $R$-Square or $r^2$ in many texts. In this example R-Square has a value of 0.997. R-square is a statistical measure that estimates the proportion of variation from the mean of the response that can be explained by the model, rather than to random error. In other words, $r^2$ quantifies the proportion of variance in the dependent response variable(s) that can be predicted from the independent variable, also called the predictor variable(s). This example indicates that there is a $(1 - 0.997237)$ or 0.276% probability that a type I error could take place, thereby erroneously concluding that the means are equal. Equivalently, this is indicating a 97.24% $(1 - \{0.276 \cdot 100\})$ confidence a Type I error has not occurred (one can conclude correctly that the means are equal, given an $\alpha$ of 0.05).
Table 3.14: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>1.4362e+13</td>
<td>1.436e+12</td>
<td>180.4515</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>3.9794e+10</td>
<td>7.9587e+9</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>15</td>
<td>1.4401e+13</td>
<td></td>
<td>&lt; .0001*</td>
</tr>
</tbody>
</table>

Table 3.15: Summary of Fit Table

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

However, the RSquare value in Table 3.15 does not adjust for the number of model parameters and, for models with a large number of parameters, RSquare is adjusted. The adjusted RSquare accounts for models with many parameters, by using degrees of freedom available in the data. Adjusted RSquare can protect against modeling with too many parameters. A model with too many parameters may be modeling noise versus the phenomenon of predictive value. The adjusted RSquare here is computed to be 0.992, which is still a good fit.

This concludes a description of how to interpret the results of the hypothesis testing and the regression analysis. The following sections report this same type of analysis for the collected smart-grid data and relevant research hypotheses and a regression test to rank the main effects and interactions in terms of relative importance.
3.10. Analysis of Each Response in the Experiment

This research work investigates what effect, if any, the above factors — namely Meters, Area, Lower-Bound and Equipment-Price, have an effect on several key characteristics of a network design, specifically Total Network Cost, Equipment Count, and Average Link Distance. Total network cost is clearly driven by the quantities of the stated factors. But the question is: by how much? And, more accurately, is the effect statistically significant? The analysis for Total Network Cost is presented first, then in subsequent sections, Equipment Count and Average Link Distance.

In the following development, experimental factors and experimental responses under study are capitalized (e.g., meters, Area, Lower-Bound, Equipment-Price, Total network cost, Equipment Count and Average Link Distance). This is to differentiate them from any general discussion of electric meters, coverage area, link distance, costs and price.

3.10.1. Total Network Cost

Knowing how factors effect the total cost of a network is important because practitioners often must report the impact of decisions on Total Network Cost when considering alternatives to the strategy, plans, or design alternatives of a network initiative. Business decisions are finalized, based in large part, on Total Network Cost.

Increasingly, business and technical cost and support objectives will include the parsimonious use of sites. Fewer sites means reduced “urban furniture” and furthering a “green” (planet-friendly) network. It is useful then to have available a tool to allow practitioners to quickly estimate the Total Network Cost. Such a tool can also allow the practitioner to build intuition about a variety of possible scenarios. This includes assessing the potential impact of changes in related factors and levels that underpin a design. Such intuition is particularly valuable for operating in a rapidly changing field such as smart-grid.
3.10.1.1. Single-Factor Analysis

The following section performs single-factor ANOVA on experimental observations to explore the effect, if any, of each of four factors on the Total Network Cost response. The four factors are the number of meters, the Area covered, the Lower-Bound on the link traffic, and the Equipment-Price.

3.10.1.1.1 The Influence of Meters on Total Network Cost

From a practical perspective, meters is a key determinant of how much network infrastructure is required to service these meters. The quantity of meters is expected to be a significant driver of the Total Network Cost of a program. The expectation is that as the level of meters is increased, the cost of the network increases.

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Meters has no relationship with the Total Network Cost. Thus, the hypothesis tested and its alternative are:

Experiment 1:

$H_0$ #1: The average Total Network Cost ($Y_1$) is the same for both levels of Meters ($X_1$).

$H_1$ #1: The average Total Network Cost ($Y_1$) is not the same for both levels of Meters ($X_1$).

![Figure 3.6: LS Means Plot of Meters vs. Total Network Cost](image)
Figure 3.6 depicts two levels of Meters on the abscissa (5,000 and 10,000), and the resultant average Total Network Cost on the ordinate of the plot. One can observe the slope of the line connecting the two costs. If the line connecting the two costs were flat, clearly there would be no difference in Total Network Cost as the level of Meters is varied. But to the degree the line has a slope, is to the degree there is a difference in Total Network Cost as the level of Meters change. So in this experiment, the number of Meters has two levels, and the mean Total Network Cost for each tested level is shown in what is called the least square means plot in Figure 3.6.

A one-factor analysis of variance, in Table 3.16, shows that the Meter levels have significantly different Total Network Costs, given a threshold of $\alpha = 0.05$ and a $p$-value equal to 0.0002. This $p$-value is indicated in the table as “Prob > F,” signifying that for an $F$-ratio value of 101, the probability is less than 0.0002 that the means of the response between the two different factor levels are equal. Thus, the null hypothesis $H_0$ #1 (the means are equal) is rejected and we can assume that Meters level has a statistically significant association with respect to Total Network Cost.

Table 3.16: ANOVA for Meters

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters</td>
<td>1</td>
<td>1</td>
<td>8.031E+11</td>
<td>101</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

As further evidence, a $t$-test is performed which determines if the two sets of data, at the two levels of the factor, are significantly different from each other. It uses the test statistic that follows a $t$-distribution. The two sets of data shown in Table 3.17 is the least-squares-means Student’s $t$-test comparing the means at the 10,000 versus 5,000 meter levels, and gives the result, for a two-tailed test, of $Prob > |t| = 0.0002$. The asterisk notes that this value is statistically significant (less than the $\alpha = 0.05$ threshold). In the table’s graphic, the sampling Student’s $t$ distribution curve is shown relative to the marker on the abscissa marking the computed $t$-ratio. The
large magnitude difference between the marker’s value and zero illustrates how large the \( t \)-ratio is, which confirms how extremely unequal the observed means are.

Table 3.17: Student’s \( t \)-test for Meters Comparing 10,000 with 5,000

<table>
<thead>
<tr>
<th>Difference</th>
<th>( 448,081 )</th>
<th>( t ) Ratio</th>
<th>( 10.04535 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>( 44,606 )</td>
<td>DF</td>
<td>( 5 )</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>( 562,743 )</td>
<td>( Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>( 333,418 )</td>
<td>( Prob &gt; t )</td>
<td>(&lt;.0001^* )</td>
</tr>
<tr>
<td>Confidence</td>
<td>( 0.95 )</td>
<td>( Prob &lt; t )</td>
<td>( 0.9999 )</td>
</tr>
</tbody>
</table>

The letters report, shown in Table 3.18, assigns the observations for each level of Meters into statistically separate groups indicating visually that, based on the sample variance, the group Total Network Cost means are significantly different from each other. The 10,000-meter Total Network Cost mean falls in group A and the 5000-meter Total Network Cost mean falls in a different group B.

Table 3.18: Letters Report Meters

<table>
<thead>
<tr>
<th>Meters</th>
<th>( t )-test Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>A</td>
<td>( 1,634,354 )</td>
</tr>
<tr>
<td>5000</td>
<td>B</td>
<td>( 1,186,274 )</td>
</tr>
</tbody>
</table>

This verifies that the level of Meters has a statistically significant effect on the Total Network Cost. As the number of Meters increase, there is a concomitant increase in infrastructure needed to connect and transport the increased traffic across an increased area. In this case, doubling the number of Meters amounts to a 38% increase in Total Network Cost (from Table 3.18 \((1.63 - 1.18)/1.18\)). Despite changes across all factors and their levels, there is an observed increase in Total Network Cost due to changes in Meters.
3.10.1.1.2 The Influence of Area on Total Network Cost

The area covered in a deployment is an important factor in the total cost of a network. In practical networks, the need to span larger areas requires equipment that can reach greater distances, and is generally more costly. The Average Link Distance generally increases as area increases. Thus increased distances typically result in more costly equipment, and translates intuitively into a more costly Total Network Cost. In this experiment, the expectation is that as the level of Area is increased, the cost of the network increases.

For formal experimentation purposes, again, the opposite conjecture is statistically tested: the level of Area makes no difference regarding the Total Network Cost. Thus, the hypothesis tested and its alternative are:

Experiment 2:

$H_0 \#2$: The average Total Network Cost ($Y_1$) is the same for both levels of Area ($X_2$).

$H_1 \#2$: The average Total Network Cost ($Y_1$) is not the same for both levels of Area ($X_2$).

Area has two levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.7. Note the slope of the line due to different mean costs for the two levels, hinting at a likelihood that a change in Area has an effect on Total Network Cost. The following statistical analysis shall confirm this initial impression.
A one-factor analysis of variance shown in Table 3.19, formally confirms that Area levels do have significantly different average Total Network Cost given a threshold of $\alpha = 0.05$, and a $p$-value equal to $< 0.0001$. Thus, the null hypothesis $H_0$ #2 (the means are equal) is rejected and we can assume that Area level has a statistically significant association with respect to Total Network Cost.

Table 3.19: ANOVA for Area

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>1</td>
<td>1</td>
<td>3.0619E+12</td>
<td>385</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

As further evidence, shown in Table 3.20, the least-squares-means Student’s $t$-test compares the means at the Area levels of 500 versus 100. This table shows the result for a two-tailed test of $\text{Prob} > |t| = < 0.0001$. The sampling Student’s $t$ distribution curve is shown relative to the marker on the abscissa marking the computed $t$-ratio. The large magnitude difference between the marker’s value and zero illustrates how large the $t$-ratio is, which confirms how extremely unequal the observed means are.
Table 3.20: Student’s $t$-test for Area Comparing 500 with 100 Areas

<table>
<thead>
<tr>
<th>Difference</th>
<th>874,911</th>
<th>$t$ Ratio</th>
<th>19.61429</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>44,606</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>989,574</td>
<td>$Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>760,248</td>
<td>$Prob &gt; t&lt; .0001*</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>$Prob &lt; t&lt; 1.0000</td>
<td></td>
</tr>
</tbody>
</table>

The letters report, shown in Table 3.21, assigns the observations for each level of Area into statistically separate groups indicating that, based on the sample variance, the group Total Network Cost means are significantly different from each other.

Table 3.21: Letters Report Area

<table>
<thead>
<tr>
<th>Area</th>
<th>$t$-test Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>A</td>
<td>1,847,769</td>
</tr>
<tr>
<td>100</td>
<td>B</td>
<td>972,858</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Area has a statistically significant effect on the Total Network Cost. As the area increases, there is a concomitant increase in infrastructure needed to connect meters and infrastructure equipment. In this case, increasing the Area five-fold, amounts to a 90% (computed from the Total Network Costs in the table) increase in Total Network Cost, across all changes in numbers of Meters, Lower-Bound, and Equipment cost.

3.10.1.1.3 The Influence of Lower-Bound on Total Network Cost

In the design of a network, practitioners have the ability to set a lower bound on the amount of traffic that is allowed to flow through equipment. By placing a lower bound on equipment, equipment with low traffic are disallowed while other equipment
and links are augmented to compensate, thus potentially reducing the total number of sites, equipment and links in a design. Intuitively, the more stringent (higher) the lower bound, the more equipment (and sites and potentially links) can be eliminated from the network. In this experiment, the conjecture is that, as the level of Lower-Bound is changed, the cost of the network changes.

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Lower-Bound makes no difference regarding the Total Network Cost. Thus, the hypothesis tested and its alternative are:

**Experiment 3:**

\[
H_0 \; 3: \text{The average Total Network Cost (} Y_1 \text{) is the same for both levels of Lower-Bound (} X_3 \text{).}
\]

\[
H_1 \; 3: \text{The average Total Network Cost (} Y_1 \text{) is not the same for both levels of Lower-Bound (} X_3 \text{).}
\]

The Lower-Bound has two levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.8. Note the lack of significant slope in the plot, hinting at a conclusion that there is no difference in the means of Total Network Costs at the two levels of Lower-Bound.

![Figure 3.8: LS Means Plot Lower-Bound vs. Total Network Cost](image)

Figure 3.8: LS Means Plot Lower-Bound vs. Total Network Cost
A one-factor analysis of variance, in Table 3.22, confirms that the Lower-Bound levels do not have significantly different Total Network Costs, given a threshold of \( \alpha = 0.05 \), and a \( p \)-value equal to \(< 0.4335 \). Thus, the null hypothesis \( H_0 \) (the means are equal) is not rejected and we can assume that Lower-Bound level does not have a statistically significant association with respect to Total Network Cost.

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>5,766,314,064</td>
<td>1</td>
<td>0.4335</td>
</tr>
</tbody>
</table>

Shown in Table 3.23, the least-squares-means Student’s \( t \)-test compares the means at the 20 versus 200 Lower-Bound levels, and gives the result, for a two-tailed test, of \( \text{Prob} > |t| = 0.4335 \). In this graphic, the sampling student’s \( t \) distribution curve is shown, and the line indicating the computed \( t \)-ratio is drawn at a point close to the center. This area illustrates how likely the probability is that the observed means are equal, thus leading to a high \( p \)-value (higher than the significance threshold of \( \alpha = 0.5 \) indicating a probability that the two observed means are equal).

<table>
<thead>
<tr>
<th>Difference</th>
<th>Std Err Dif</th>
<th>( t ) Ratio</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37968</td>
<td>0.851193</td>
<td></td>
</tr>
<tr>
<td>Std Err Dif</td>
<td>44,606</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>152,631</td>
<td>( \text{Prob} &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>-76,695</td>
<td>( \text{Prob} &gt; t )</td>
<td>0.2168</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>( \text{Prob} &lt; t )</td>
<td>0.7832</td>
</tr>
</tbody>
</table>
The letters report, shown in Table 3.24, assigns the observations for each level of Lower-Bound into a statistically similar group indicating that, based on the sample variance, the group Total Network Cost means are not significantly different from each other.

Table 3.24: Letters Report Lower-Bound

<table>
<thead>
<tr>
<th>Lower-Bound</th>
<th>t-test Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>A</td>
<td>1,429,298</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>1,391,330</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Lower-Bound does not have a statistically significant effect on the Total Network Cost. As the Lower-Bound changes, there is no statistically significant change in infrastructure needed for the network. In this case, increasing the Lower-Bound by an order of magnitude accounts for only an increase of 3% in Total Network Cost (computed from the values in Table 3.24). Thus, a counterintuitive result is obtained for the conjecture that a Lower-Bound on equipment has an effect on (reducing) Total Network Cost.

3.10.1.1.4 The Influence of Equipment-Price on Total Network Cost

From a practical perspective, and intuitively obvious, the Equipment-Price should directly drive the Total Network Cost. It is expected that as the Equipment-Price increases, so does the Total Network Cost. In this section we quantify this conjecture by testing an increase of Equipment-Price ten-fold.

For formal experimentation purposes, the following conjecture is statistically tested, that the level of Equipment-Price makes no difference regarding the Total Network Cost. Thus, the hypothesis tested is:
Experiment 4:

$H_0$ #4: The average Total Network Cost ($Y_1$) is the same for both levels of Equipment-Price ($X_4$).

$H_1$ #4: The average Total Network Cost ($Y_1$) is not the same for both levels of Equipment-Price ($X_4$).

The Equipment-Price has two levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.9. Note the significant slope of the characteristic.

![Figure 3.9: LS Means Plot Equipment-Price vs. Total Network Cost](image)

A one-factor analysis of variance, in Table 3.25, shows that the Equipment-Price levels have significantly different Total Network Costs, given a threshold of $\alpha= 0.05$, and a $p$-value equal to $> 0.0001$. Thus, the null hypothesis $H_0$ #4 (the means are equal) is rejected and we can assume that Equipment-Price level has a statistically significant association with respect to Total Network Cost.

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>9.7252E+12</td>
<td>1,222</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>
The least-squares-means Student’s $t$-test, shown in Table 3.26 compares the means at the 2,000 versus 200,000 levels, and gives the result, for a two-tailed test, of $Prob > |t| \geq 0.0001$. The graphic illustrates how extreme the $t$-ratio value is from the mean of the $t$-distribution.

Table 3.26: Student’s $t$-test for Equipment-Price Comparing 200,000 with 20,000 Prices

<table>
<thead>
<tr>
<th>Difference</th>
<th>1,559,261</th>
<th>$t$ Ratio</th>
<th>34.95648</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>44,606</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>1,673,924</td>
<td>$Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>1,444,599</td>
<td>$Prob &gt; t &lt; .0001^*$</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>$Prob &lt; t$</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The letters report, shown in Table 3.27, indicates that, based on the sample variance, the group Total Network Cost means are significantly different from each other.

Table 3.27: Letters Report Meters

<table>
<thead>
<tr>
<th>Equipment-Price</th>
<th>$t$-test Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>200,000</td>
<td>A</td>
<td>2,189,945</td>
</tr>
<tr>
<td>20,000</td>
<td>B</td>
<td>630,683</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Equipment-Price has a significant effect on the Total Network Cost, as expected. In this case, by increasing the cost of each unit of equipment by an order of magnitude, from the table, the Total Network Cost increases by 247% ($2,189,945 - 630,683/630,682$). It can be concluded, now for this scenario quantitatively, that the price of equipment is an important factor when planning networks, and often drives competition among vendors to offer lower-priced
equipment. By doing so, vendors still must meet technical requirements, but their pricing has a significant effect on the overall network cost.

3.10.1.2. Pair-wise Analysis

Pair-wise analysis, also called *post hoc* analysis, is next performed on pairs of factors to determine if any two factors, when combined, significantly influence the response Total Network Cost. If so, they are deemed an interaction, which sheds light on which factor pairs effect the response. If there are interactions among the factors tested, this type of pair-wise analysis uncovers the specific combination(s) that have a statistically significant effect on the response.

3.10.1.2.1 The Influence of Meters*Area on Total Network Cost

From a practical perspective, the number of Meters when combined with the service Area to be covered, characterizes a density, notionally, meters-per-area. Density is an interesting characteristic of a network deployment, since it is typically wide-ranging, from a dense urban setting to a sparse rural or remote setting. The number of meters per square mile is used here as the unit of measure for meter density. Intuitively one expects a more dense deployment to be more significant in driving Total Network Cost than a sparse deployment. This is because, to a certain point, higher density drives heftier equipment capacity needs, while larger Area drives only the less costly higher equipment range capability. This however is not altogether straightforward. Extremes in density, either way, tend to drive up the Total Network Cost, for example a very large number of meters in a small area, or for a scarcity of meters across a large Area, both would intuitively tend to drive up Total Network Cost.

For formal experimentation purposes, however, we state the following conjecture: all combinations of Meters and Area (denoted Meters*Area) make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:
Experiment 5:

$H_0\ 5$: The average Total Network Cost ($Y_1$) is the same for all levels of all combinations of Meters*Area ($X_5$).

$H_1\ 5$: The average Total Network Cost ($Y_1$) is not the same for all levels of all combinations of Meters*Area ($X_5$).

The Meters*Area interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.10. For Area, the slope, when Meters are fixed at 5,000 (the lower line), is significant, indicating the Area*Meters interaction term influences Total Network Cost. For Meters, the vertical distance between endpoints at Area = 100, and the vertical distance between the endpoints at Area = 500, being greater than 0 (there exists a gap), again indicates that Meters has an effect on Total Network Cost about both values of Area.

Figure 3.10: LS Means Plot Meters*Area Interaction vs. Total Network Cost

A two-factor analysis of variance, in Table 3.28, proves that the Meter*Area levels have significantly different Total Network Costs, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.0006. Thus, the null hypothesis $H_0 \neq 5$ (the means are equal) is rejected and we can assume that Meters*Area level has a statistically significant association with respect to Total Network Cost.
Now that there are more than two means to test a Tukey HSD test can be used to compare all four means at once. A Tukey HSD test compares the means at the four different combination levels, and gives the result, in a letters report, assigning the observations for each level of Meters*Area into statistically separate groups. Shown in Table 3.29, based on the sample variance, the group Total Network Cost means are all significantly different from each other.

Thus it is shown that the levels of Meters*Area, also representing levels of density, have a statistically significant effect on the Total Network Cost at each level. As the number of Meters*Area increase, Total Network Cost is driven in major increments by Area, and minor increments within Area, by Meters. Each increment from lowest to highest increases Total Network Cost incrementally from the lowest by 29%, 39% and 43% respectively, computed from the Total Network Costs in Table 3.29. Because this interaction is significant, this gives support to the fact that the two individual factors are also significant (as shown in the prior section).
These results are interesting in that there is a statistically significant difference at each density level, the largest increase coming as the average Area is increased five-fold, then within that, as Meters are doubled. The practical implication is that both factors play a significant role in the Total Network Cost, individually and jointly, with Area being the major driver for this set of tests.

3.10.1.2.2 The Influence of Meters*Lower-Bound on Total Network Cost

The two-factor interaction of Meters and Lower-Bound, (labeled as Meters*Lower-Bound) do not give as intuitive a picture, feeling or “ring” as does the preceding interaction that formed the density concept. The two-factor interaction of Meters*Lower-Bound on Total Network Cost gives only an abstract intuitive notion, namely how might the Total Network Cost have an effect between two different types of of equipment, differing in Lower-Bound, used in two different levels of meter deployments. This does not give an obvious practical interpretation.

However, for formal experimentation purposes, the following conjecture is statistically tested, namely, all combinations of Meters*Lower-Bound make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 6:

\[ H_0 \#6: \text{The average Total Network Cost (} Y_1 \text{) is the same for all levels of all combinations of Meters*Lower-Bound (} X_6 \text{).} \]

\[ H_1 \#6: \text{The average Total Network Cost (} Y_1 \text{) is not the same for all levels of all combinations of Meters*Lower-Bound (} X_6 \text{).} \]

The Meters*Lower-Bound, interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.11. Note the minor degree of slopes, indicating not much difference in terms of Lower-Bound, and note the vertical distance between points, indicating their likely is some difference between
Meters levels.

Figure 3.11: LS Means Plot Meters*Lower-Bound, vs. Total Network Cost

A two-factor analysis of variance, in Table 3.30, shows that the Meters*Lower-Bound levels do not have significantly different Total Network Costs, given a threshold of $\alpha = 0.05$, and a p-value equal to 0.2244. Thus, the null hypothesis $H_0$ (the means are equal) is not rejected and we can assume that Meters*Lower-Bound level does not have a statistically significant association with respect to Total Network Cost.

Table 3.30: Two-Factor ANOVA for Meters*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>15,285,000,000</td>
<td>2</td>
<td>0.2244</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Meters*Lower-Bound do not have a statistically significant effect on the Total Network Cost.

While it can be speculated that equipment on both ends of a link with a more stringent Lower-Bound on traffic would be expected to lower Total Network Cost, by eliminating some low traffic links, the evidence, for this set of data, do not bear that out. Equipment with a higher level of Lower-Bound, for this data, does not lower Total Network Cost.
The Influence of Meters*Equipment-Price on Total Network Cost

For the interaction term Meters*Equipment-Price, again there is no readily relatable practical notion for the combination. It would seem obvious that as both factors, Meters and Equipment-Price are increased individually or jointly, the Total Network Cost would increase.

For formal experimentation purposes, the opposite is tested, that all combinations of Meters and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 7:

\[ H_0 \#7: \text{The average Total Network Cost} (Y_1) \text{ is the same for all levels of all combinations of Meters*Equipment-Price} (X_7). \]

\[ H_1 \#7: \text{The average Total Network Cost} (Y_1) \text{ is not the same for all levels of all combinations of Meters*Equipment-Price} (X_7). \]

The Meters*Equipment-Price interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.12. Note the significant slope, and the vertical distance gap between points. This points to the expected strong correlation between Equipment Price and Total Network Cost, and some correlation between Meters and Total Network Cost.
Figure 3.12: LS Means Plot Meters*Equipment-Price Interaction vs. Total Network Cost

A two-factor analysis of variance, in Table 3.31, shows that the Meters*Equipment-Price levels have significantly different Total Network Costs, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.0227. Thus, the null hypothesis $H_0$ (#7 (the means are equal) is rejected and we can assume that Meters*Equipment-Price interaction is statistically significant with respect to Total Network Cost.

Table 3.31: Two-Factor ANOVA for Meters*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>84047000000</td>
<td>11</td>
<td>0.0227</td>
</tr>
</tbody>
</table>

As further evidence, the Tukey HSD test compares the means at the four different combination levels, and gives the result, in a letters report, assigning the observations for each level of Meters*Equipment-Price into statistically separate groups. Shown in Table 3.32, based on the sample variance, the group Total Network Cost means are all significantly different from each other.
Thus it is shown that the levels of Meters*Equipment-Price, have a statistically significant effect on the Total Network Cost at each level. As the number of Meters*Equipment-Price increase, Total Network Cost is driven in major increments by Equipment-Price and minor increments within Equipment-Price by Meters. Each increment from lowest to highest increases mean Total Network Cost incrementally from the lowest (baseline at 0%) by 63%, 142%, and 31% respectively (computed from the Total Network Costs in the table). Given that this interaction is significant, this gives support to the fact that the two individual factors are significant.

These results are intuitively obvious, in that the expectation is met, that there is a statistically significant difference at each increase in primarily price, and Meters secondarily. As price is increased 10-fold, Total Network Cost increases by 142%. This underscores the importance of competition of vendors on Equipment-Price when designing networks, particularly as the size of the project grows in terms of Meters, in order to achieve a competitive Total Network Cost.

3.10.1.2.4 The Influence of Area*Lower-Bound on Total Network Cost

The two-factor interaction of Area and Lower-Bound (Area*Lower-Bound) also do not give as intuitive an interpretation as do some preceding interactions. Given this, there is no intuitive expectation for this variable. One can imagine that the combination behaves as the individuals do, such that, as both Area and Lower-Bound increase,
Total Network Cost increases.

For formal experimentation purposes, the opposite conjecture is statistically tested. All combinations of Area*Lower-Bound make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 8:

\( H_0 \ #8: \) The average Total Network Cost \((Y_1)\) is the same for all levels of all combinations of Area*Lower-Bound \((X_8)\).

\( H_1 \ #8: \) The average Total Network Cost \((Y_1)\) is not the same for all levels of all combinations of Area*Lower-Bound \((X_8)\).

The Area*Lower-Bound interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.13. Note the wide gap in vertical distance between different Area levels. There is not much slope, however, between Lower-Bound levels. So this points to the expectation that Meters likely is influential but Lower-Bound not as influential in terms of an effect on Total Network Cost. The analysis will determine if the Meters influence pushes the total interaction term across the threshold for overall statistical significance of this interaction term.

![Figure 3.13: LS Means Plot Area*Lower-Bound vs. Total Network Cost](image)

Figure 3.13: LS Means Plot Area*Lower-Bound vs. Total Network Cost
A two-factor analysis of variance, in Table 3.33, shows that the Area*Lower-Bound levels do not have significantly different Total Network Costs, given a threshold of \( \alpha = 0.05 \), and a \( p \)-value equal to 0.4175. Thus, the null hypothesis \( H_0 \) (the means are equal) is not rejected and we can assume that Area*Lower-Bound level does not have a statistically significant association with respect to Total Network Cost.

Table 3.33: Two-Factor ANOVA for Area*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area*Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>6208849014</td>
<td>1</td>
<td>0.4175</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Area*Lower-Bound do not have a statistically significant effect on the Total Network Cost.

3.10.1.2.5 The Influence of Area*Equipment-Price on Total Network Cost

The interaction term Area*Equipment-Price has no intuitive practical notion as a combination of these two factors. The expectation however is that as individual or both factors are increased, so would Total Network Cost.

For the formal experimentation the opposite conjecture is statistically tested. All combinations of Area and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 9:

\( H_0 \) #9: The average Total Network Cost \((Y_1)\) is the same for all levels of all combinations of Area*Equipment-Price \((X_9)\).

\( H_1 \) #9: The average Total Network Cost \((Y_1)\) is not the same for all levels of all combinations of Area*Equipment-Price \((X_9)\).
The Area*Equipment-Price interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.14. Note the strong slope and the large gap between vertical points, hinting at the likelihood of both Equipment Price and Area have an effect on Total Network Cost.

Figure 3.14: LS Means Plot Area*Equipment-Price Interaction vs. Total Network Cost

A two-factor analysis of variance, in Table 3.34, shows that the Area*Equipment-Price levels have significantly different Total Network Costs, given a threshold of $\alpha = 0.05$, and a p-value equal to 0.0005. Thus, the null hypothesis $H_0$ (the means are equal) is rejected and we can assume that Area*Equipment-Price interaction is statistically significant with respect to Total Network Cost.

Table 3.34: Two-Factor ANOVA for Area*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>4.9133E+11</td>
<td>62</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

The Tukey HSD test compares the means at the four different combination levels, and gives the result, in the letters report, assigning the observations for each level of Area*Equipment-Price into statistically separate groups. Shown in Table 3.35, based
on the sample variance, the group Total Network Cost means are all significantly different from each other.

Table 3.35: Letters Report Area*Equipment-Price

<table>
<thead>
<tr>
<th>Area*Equipment-Price</th>
<th>Tukey HSD Group</th>
<th>Avg. Total Network Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 200000</td>
<td>A</td>
<td>2,802,637</td>
</tr>
<tr>
<td>100 200000</td>
<td>B</td>
<td>1,577,252</td>
</tr>
<tr>
<td>500 20000</td>
<td>C</td>
<td>892,902</td>
</tr>
<tr>
<td>100 20000</td>
<td>D</td>
<td>368,464</td>
</tr>
</tbody>
</table>

Thus it is shown that the levels of Area*Equipment-Price, have a statistically significant effect on the Total Network Cost at each level. As the interaction combination Area*Equipment-Price increases, Total Network Cost is driven strongly for the first increment (Area) 142%, then 77% and 78%. Given that this interaction is significant, this gives support to the fact that the two individual factors are significant, as shown in the single factor tests in the prior section.

3.10.1.2.6 The Influence of Lower-Bound*Equipment-Price on Total Network Cost

The two-factor interaction of Lower-Bound and Equipment-Price again does not give an intuitive notion when taken together. Given this, there is no intuitive expectation for the combination, other than the combination may increase the Total Network Cost for increases in Equipment Price and decrease the Total Network Cost for increases in Lower-Bound.

For the experiment, the opposite conjecture is statistically tested. All combinations of Lower-Bound*Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:
Experiment 10:

\( H_0 \#10: \) The average Total Network Cost \((Y_1)\) is the same for all levels of all combinations of Lower-Bound*Equipment-Price \((X_{10})\).

\( H_1 \#10: \) The average Total Network Cost \((Y_1)\) is not the same for all levels of all combinations of Lower-Bound*Equipment-Price \((X_{10})\).

The Lower-Bound*Equipment-Price interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.15. The strong slope for Equipment Price bodes well that it influences Total Network Cost, yet the lack of any gap in vertical distance between points indicate little effect on Total Network Cost by increases in Lower-Bound.

![Figure 3.15: LS Means Plot Lower-Bound*Equipment-Price vs. Total Network Cost](image)

A two-factor analysis of variance, in Table 3.36, shows that the Lower-Bound*Equipment-Price levels do not have significantly different Total Network Costs, given a threshold of \( \alpha = 0.05 \), and a \( p \)-value equal to 0.6111. Thus, the null hypothesis \( H_0 \#10 \) (the means are equal) is rejected and we can assume that Lower-Bound*Equipment-Price level does not have a statistically significant association with respect to Total Network Cost.
Table 3.36: Two-Factor ANOVA for Lower-Bound*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>NP</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>2337891728</td>
<td>0</td>
<td>0.6111</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Lower-Bound*Equipment-Price do not have a statistically significant effect on the Total Network Cost.

3.10.1.3. Total Network Cost Regression Analysis

This section presents the results of performing a least squares fit of the four main effects and six interaction terms to derive a prediction equation for Total Network Cost (Total Network Cost). The results of the regression calculations are shown in Table 3.37. The table lists the independent \( x \) term (for example in the first row: \( x = \) Equipment-Price), the estimate of the \( \beta_i \) coefficient (e.g., 8.6625632), the standard error, the computed \( t \)-Ratio, and the \( p \)-value ("\( \text{Prob} > |t| \")"). The table is sorted by \( t \)-ratio value, which highlights the variables that have the most effect on the response.

The Analysis of Variance Table 3.38 shows a good fit of the regression line to the data, with a \( p \)-value (\( \text{Prob} > F \)) of \(< 0.0001^* \) (the asterisk denoting statistical significance). The Summary of Fit Table 3.39 shows the coefficient of determination, also called “RSquare,” has a value of 0.997. This is indicating that \( (1 - 0.997237 =) 0.276\% \) of the time a type I error could take place, of concluding, in error, that the means are equal. Equivalently, this is indicating a 97.24\% \( (1 - \{0.276 \cdot 100\} =) \) confidence a Type I error has not occurred (thus concluding correctly the means are equal).
Table 3.37: Sorted Parameter Estimates

| Term                              | Estimate $\beta_i$ | Std Error | $t$-Ratio | Prob $>|t|$ |
|-----------------------------------|--------------------|-----------|-----------|------------|
| Equipment-Price                   | 8.6625632          | 0.24781   | 34.96     | <.0001*    |
| Area                              | 2187.2778          | 111.5145  | 19.61     | <.0001*    |
| Meters                            | 89.616125          | 8.921158  | 10.05     | 0.0002*    |
| (Area-300)*                       | 0.0097354          | 0.001239  | 7.86      | 0.0005*    |
| (Equipment-Price-110000)          |                    |           |           |            |
| (Meters-7500)*(Area-300)          | 0.2040006          | 0.044606  | 4.57      | 0.0060*    |
| (Meters-7500)*                    | 0.0003221          | 9.912e−5  | 3.25      | 0.0227*    |
| (Equipment-Price-110000)          |                    |           |           |            |
| (Meters-7500)*(Lower-Bound-110)   | −0.137369          | 0.099124  | −1.39     | 0.2244     |
| (Area-300)*(Lower-Bound-110)      | 1.0943924          | 1.23905   | 0.88      | 0.4175     |
| Lower-Bound                       | 210.93403          | 247.81    | 0.85      | 0.4335     |
| (Lower-Bound-110)*                | 0.0014923          | 0.002753  | 0.54      | 0.6111     |
| (Equipment-Price-110000)          |                    |           |           |            |
Table 3.38: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>1.4362e+13</td>
<td>1.436e+12</td>
<td>180.4515</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>3.9794e+10</td>
<td>7.9587e+9</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>15</td>
<td>1.4401e+13</td>
<td></td>
<td>&lt; .0001*</td>
</tr>
</tbody>
</table>

Table 3.39: Summary of Fit Table

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

This RSquare value in Table 3.39 however does not adjust for the number of parameters in the model, and for models with a large number of parameters, further testing is prudent. The adjusted RSquare accounts for models with many parameters, by using degrees of freedom available in the data. Adjusted RSquare can protect against modeling with too many parameters. A model with too many parameters may be modeling noise versus the phenomenon of predictive value. The adjusted RSquare here is computed to be 0.992, which is still an excellent fit. The results in Table 3.37 show the terms most useful in predicting Total Network Cost. By being equal to or less
than an alpha of 0.05 (also shown by the asterisk), the following terms are significant: Equipment-Price ($p$-value < .0001), Area (> .0001), Meters (0.0002), and the interaction between Area*Equipment-Price, Meters*Area, and Meters*Equipment-Price. The Lower-Bound is not statistically significant as a main effect nor in interaction terms.

Figure 3.16: Actual by Predicted Plot

Figure 3.16 shows actual Total Network Cost data points plotted against predicted data points using the regression equation of Figure 3.17. The 95% confidence dotted lines closely straddle the sloping regression line, indicating a good confidence of fit. These curves do not straddle the dotted horizontal mean line, thereby further indicating the observed data is far from merely reflecting random data about the mean.

The actual prediction expression for Total Network Cost, is shown in Figure 3.17. This equation can be constructed from the analysis output.
Note that centering and scaling \[62\] are used to offset the coefficients in the equation of Figure 3.17. Scaling the variances to have a mean of zero and standard deviation of one allows changes to be standardized across variables, both independent and dependent. For example, a change of one standard deviation in Meters becomes approximately equivalent to a change of one standard deviation in Area. Thus, scaling
puts all dependent responses and independent predictors on par, with respect to their variations. By centering, or subtracting off the variables’ medians, successive factors can be ordered in terms of the variation they explain. Without centering, both mean and variance would complicate the ability to sort factors’ relative influence. Thus scaling and centering allows for proper comparison and ranking of the importance of effects on responses.

3.10.1.4. Summary for Total Network Cost

The summary of experimental hypothesis testing is shown in Table 3.40. When viewing all tests holistically, it is recommended practice to look first at the interactions and then to the main effects. Therefore, important interactions include Area*Equipment-Price, Meters*Equipment-Price, and Meters*Area. The important main effects are Meters, Area, and Equipment-Price. Note the important interactions support and strengthen the conclusion of the importance of the main effects. Thus, for this set of solution instances, Meters, Area, and Equipment-Price are significant factors that have an effect on Total Network Cost.

Practitioners are well aware that the total cost of a network is driven by the quantity of meters, service area, and price of equipment. What is not as obvious is the degree to which a lower bound has an effect on Total Network Cost. At least for this set of instances, setting a Lower-Bound on equipment throughput is not statistically significant in driving Total Network Cost.

3.10.2. Equipment Count

When designing a communications network, the amount of infrastructure needed includes items such as base stations, access points, towers, antennas, and a site for placement of this equipment. In the present work, these sites are where repeaters and collectors are placed. The repeater or collector equipment must be situated on a pole, radio tower, or at a substation site.
Table 3.40: Summary Results for Total Network Cost

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ((H_0))</th>
<th>(Prob &gt; t)</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total Network Costs are equal for Meters</td>
<td>0.0002</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>2</td>
<td>Total Network Costs are equal for Area</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>3</td>
<td>Total Network Costs are equal for L.Bound</td>
<td>0.4335</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>4</td>
<td>Total Network Costs are equal for E.Price</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>5</td>
<td>Total Network Costs are equal for Meters* Area</td>
<td>0.0060</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>6</td>
<td>Total Network Costs are equal for Meters* L.Bound</td>
<td>0.2244</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>7</td>
<td>Total Network Costs are equal for Meters* E.Price</td>
<td>0.0227</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>8</td>
<td>Total Network Costs are equal for Area* L.Bound</td>
<td>0.4175</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>9</td>
<td>Total Network Costs are equal for Area* E.Price</td>
<td>0.0005</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>10</td>
<td>Total Network Costs are equal for L.Bound* E.Price</td>
<td>0.6111</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>
The factor Equipment Count is the total number of repeaters and collectors needed in a design. One figure of merit of a network design is site count. Typically, the fewer the sites required for a design the better. The outcome of a design’s Equipment Count may be influenced by initial conditions and designer decisions at the start of a design. The following deals with the association of these design input factors on the response Equipment Count.

3.10.2.1. Single-Factor Analysis

The following section performs single-factor ANOVA on experimental observations to explore the effect, if any, of each of four factors on the Equipment Count response. The four factors are the same as above, the number of Meters, the Area covered, the Lower-Bound on link traffic, and the Equipment-Price.

3.10.2.1.1 The Influence of Meters on Equipment Count

From a practical perspective, the number of Meters served would intuitively drive the amount of network equipment needed to transport the meter traffic. The conjectural expectation therefore is that, as the level of Meters is increased, the Equipment Count increases.

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Meters has no relationship with Equipment Count. Thus, the hypothesis tested is:

Experiment 11:

\[ H_0 \#11: \text{The average Equipment Count (} Y_2 \text{) is the same for both levels of Meters (} X_1 \text{).} \]

\[ H_0 \#11: \text{The average Equipment Count (} Y_2 \text{) is not the same for both levels of Meters (} X_1 \text{).} \]
In this study, the number of Meters has two levels, and the mean Equipment Count for each tested level is shown in the least square means plot of Figure 3.18. Note the slope would suggest there is a difference in the means, leading to a rejection of $H_0$; meaning there is a suggestion of an effect of Meters on Equipment Count.

A one-factor analysis of variance, in Table 3.41, shows that the Meter levels have significantly different Equipment Counts, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.0215. Thus, the null hypothesis $H_0$ #11 (the means are equal) is rejected and we can assume that Meters level has a statistically significant association with respect to Equipment Count.

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters</td>
<td>1</td>
<td>1</td>
<td>85.5625</td>
<td>11</td>
<td>0.0215</td>
</tr>
</tbody>
</table>

As further evidence shown in Table 3.42, the least-squares-means Student’s $t$-test compares the means at the 10,000 versus 5,000 meter levels, and gives the result, for a two-tailed test, of $Prob > |t| = 0.0215$. In the graphic, the sampling student’s
distribution curve is shown relative to the marker on the abscissa indicating the computed $t$-ratio. This illustrates how extremely unequal the observed means are, thus leading to a low $p$-value (the probability that the two observed means are equal), and conclusion the means are not equal.

Table 3.42: Student’s $t$-test for Meters Comparing 5000 with 10,000 Meters

<table>
<thead>
<tr>
<th>Difference</th>
<th>4.62500</th>
<th>$t$ Ratio</th>
<th>3.298841</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>1.4020</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>8.22897</td>
<td>$Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>1.02103</td>
<td>$Prob &gt; t$</td>
<td>0.0108$*</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>$Prob &lt; t$</td>
<td>0.9892</td>
</tr>
</tbody>
</table>

The letters report , shown in Table 3.43, assigns the observations for each level of Meters into statistically separate groups indicating that, based on the sample variance, the group Equipment Count means are significantly different from each other.

Table 3.43: Letters Report Meters

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Equipment Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>5000</td>
<td>B</td>
<td>6</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Meters has a statistically significant effect on the Equipment Count. As the number of Meters increase, there is an increase in infrastructure needed to connect and transport the increased traffic. In this case, doubling the number of Meters, amounts to a $((11 - 6)/6) = 78\%$ increase in Equipment Count across changes in the other factors of Area, Lower-Bound and Equipment cost.

In real networks, Equipment Count represents repeaters and collectors needed to
aggregate the meter traffic in the direction of the core network. As the number of Meters increase, the number of repeaters and collectors increase, an intuitive result.

3.10.2.1.2 The Influence of Area on Equipment Count

The area covered in a deployment is an important factor in terms of the equipment needed for the network. In practical networks, the need to span large areas requires either more hops, therefore more equipment, or equipment with longer range capability. In this experiment, the expectation is that as the level of Area is increased, the units of Equipment Count increase.

For formal experimentation purposes, again, the opposite conjecture is statistically tested: the level of Area makes no difference regarding the mean Equipment Count. Thus, the hypothesis tested is:

Experiment 12:

$H_0 \ #12$: The average Equipment Count ($Y_2$) is the same for both levels of Area ($X_2$).

$H_1 \ #12$: The average Equipment Count ($Y_2$) is not the same for both levels of Area ($X_2$).

The Area has two levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.19. Note the slope, indicating a possible effect on to Equipment Count by Area. This type of estimate however is not accurate in all cases, as will be seen shortly (there is not a statistically significant effect).
A one-factor analysis of variance shown in Table 3.44, confirms that the Area levels do not have significantly different Equipment Counts given a threshold of $\alpha = 0.05$, and a $p$-value equal to $< 0.1201$. Thus, the null hypothesis $H_0$ (the means are equal) is not rejected and we can assume that Area level does not have a statistically significant association with respect to Equipment Count.

Table 3.44: ANOVA for Area

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>1</td>
<td>1</td>
<td>27.5625</td>
<td>4</td>
<td>0.1201</td>
</tr>
</tbody>
</table>

As further evidence shown in Table 3.45, the least-squares-means Student’s $t$-test compares the means at the 500 versus 100 Area levels, and gives the result, for a two-tailed test, of $\text{Prob} > |t| \leq 0.1201$. The sampling Student’s $t$ distribution curve is shown relative to the line indicating the computed $t$-ratio. This illustrates the area under the curve representing the possibility of a Type I error. The illustration shows far too much area under the curve, thus providing not enough evidence that the means are different.
Table 3.45: Student’s \( t \)-test for Area Comparing 500 with 100 Areas

| Difference | t Ratio | Std Err Dif | DF | Upper CL Dif | Prob > \(| t |\) | Lower CL Dif | Prob > \( t \) | Confidence | Prob < \( t \) |
|------------|---------|-------------|----|-------------|----------------|--------------|--------------|-------------|------------|
| 2.6250     | 1.872315| 1.4020      | 5  | 6.2290      | 0.1201         | −0.9790      | 0.0600       | 0.95        | 0.9400     |

The letters report, shown in Table 3.46, assigns the observations for each level of Area into a single group (A) indicating that, based on the sample variance, the group Equipment Count means are not significantly different from each other.

Table 3.46: Letters Report Area

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Equipment Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>A</td>
<td>7</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Area does not have a statistically significant effect on the Equipment Count. As the Area increases, while there is a 37% increase in units of equipment, this is not statistically significant. For this sample set of instances, Area can be increased yet the effect on Equipment Count is not statistically significant. While this may be surprising to practitioners (that larger areas do not necessarily translate into more, or higher-powered equipment), the present model represents areas within a suburban zone only, and therefore modeled with equipment that does not necessarily have a distance limitation. Resulting in Area levels being positively correlated, but not statistically significant in the present model.
3.10.2.1.3 The Influence of Lower-Bound on Equipment Count

Intuitively, the more stringent (higher) the Lower-Bound, the more links (and sites) can be eliminated from the network, reducing potentially the units of equipment needed for traffic transport. Consequently, in this experiment, the conjectural expectation is that, as the level of Lower-Bound is changed, the Equipment Count changes.

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Lower-Bound makes no difference regarding the Equipment Count. Thus, the hypothesis tested is:

Experiment 13:

$H_0$ #13: The average Equipment Count ($Y_2$) is the same for both levels of Lower-Bound ($X_3$).

$H_1$ #13: The average Equipment Count ($Y_2$) is not the same for both levels of Lower-Bound ($X_3$).

The Lower-Bound has two levels, and the mean Equipment Count for each tested level is shown in the means plot of Figure 3.20. Note the negative slope indicating there is a possibility of some effect of Lower-Bound on Equipment Count, namely a negative correlation (increased Lower-Bound decreases Equipment Count, as expected). But is this statistically significant?
A one-factor analysis of variance, in Table 3.47, shows that the Lower-Bound levels do not have significantly different Total Network Costs, given a threshold of $\alpha=0.05$, and a $p$-value equal to $<0.3718$. Thus, the null hypothesis $H_0$ (the means are equal) is not rejected and we can assume that Lower-Bound level does not have a statistically significant association with respect to Total Network Cost. The means are different, but not statistically so.

Table 3.47: ANOVA for Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>$Prob &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>7.5625</td>
<td>1</td>
<td>0.3718</td>
</tr>
</tbody>
</table>

Shown in Table 3.48, the least-squares-means Student’s $t$-test compares the means at the 20 versus 200 Lower-Bound levels, and gives the result, for a two-tailed test, of $Prob > |t| = 0.3718$. In this graphic, the sampling student’s $t$ distribution curve illustrates the area representing how likely the probability is that the observed means are equal, thus leading to a high $p$-value, and insufficient evidence to reject the null hypothesis.
Table 3.48: Student’s t-test for Lower-Bound Comparing 200 with 20

|          | Difference | t Ratio  | Std Err Dif | t Ratio  | DF | Upper CL Dif  | Prob > |t| |  | Lower CL Dif  | Prob > t |  |  | Confidence | Prob < t |
|----------|------------|----------|-------------|----------|----|---------------|--------|----| |  |                 |          |  |  |           |          |
| t Ratio  | −1.3750    |          | 1.4020      |          | 5  | 2.2290        |        |  |  | −4.9790     | 0.8141    |  |  | 0.95       | 0.1859   |
|          | −0.98074   |          |             |          |    |               |        |  |  |            |           |  |  |           |          |

The t-test letters report assigns the observations for each level of Meters into statistically separate groups, shown in Table 3.49, indicating that, based on the sample variance, the Equipment Count means are not significantly different from each other.

Table 3.49: Letters Report Lower-Bound

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Equipment Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>200</td>
<td>A</td>
<td>8</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Lower-Bound does not have a statistically significant effect on the Equipment Count. As the Lower-Bound increases, there is some increase, but not a statistically significant increase in infrastructure needed for the network. In this case, increasing the Lower-Bound by an order of magnitude accounts for only a change of -18% in Equipment Count. So while there is some correlation, a counterintuitive result is obtained from this experiment, namely increasing the Lower-Bound does not significantly reduce Equipment Count, it should be noted that an decrease of 18% may be sufficiently significant from a business perspective, and business considerations may sway the use of Lower-Bound as a technique for attempting a lowering Equipment Count by some degree.
3.10.2.1.4 The Influence of Equipment-Price on Equipment Count

From a practical perspective, increasing Equipment-Price should have an effect on Equipment Count by forcing longer links versus more equipment, when feasible. Thus, it is expected that as the Equipment-Price increases, the Equipment Count is reduced. In this section we test this conjecture by evaluating the effect on Equipment Count for a ten-fold increase in Equipment-Price.

For formal experimentation purposes, the following conjecture is statistically tested, that the level of Equipment-Price makes no difference regarding the Equipment Count. Thus, the hypothesis tested is:

Experiment 14:

\[ H_0 \ #14: \text{The average Equipment Count } (Y_2) \text{ is the same for both levels of Equipment-Price (X}_4\). \]

\[ H_1 \ #14: \text{The average Equipment Count } (Y_2) \text{ is not the same for both levels of Equipment-Price (X}_4\). \]

The Equipment-Price has two levels, and the mean Equipment Count for each tested level is shown in the means plot of Figure 3.21. Note the slope suggests a correlation between Equipment Price and Equipment Count. This will not bear out, as shown next.
A one-factor analysis of variance, in Table 3.50, shows that the Equipment-Price levels do not have significantly different Equipment Counts, given a threshold of $\alpha = 0.05$, and a $p$-value equal to $> 0.151$. Thus, the null hypothesis $H_0 \ #14$ (the means are equal) is not rejected and we can assume that Equipment-Price level does not have a statistically significant association with respect to Equipment Count.

Table 3.50: ANOVA for Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>$Prob &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>22.5625</td>
<td>3</td>
<td>0.151</td>
</tr>
</tbody>
</table>

The least-squares-means Student’s $t$-test, shown in Table 3.51 compares the means at the 2,000 versus 200,000 levels, and gives the result, for a two-tailed test, of $Prob > \left| t \right| \geq 0.151$. The graphic illustrates the critical regions representing the probability of a Type I error, and likelihood of concluding incorrectly that the means are equal.
Table 3.51: Student’s t-test for Equipment-Price Comparing 200000 with 20000 Prices

<table>
<thead>
<tr>
<th>Difference</th>
<th>2.3750</th>
<th>t Ratio</th>
<th>1.694</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>1.4020</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>1.2290</td>
<td>Prob &gt;</td>
<td>0.1510</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>−5.9790</td>
<td>Prob &gt; t</td>
<td>0.9245</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>Prob &lt; t</td>
<td>0.0755</td>
</tr>
</tbody>
</table>

Lastly, a t-test letters report, shown in Table 3.52, indicates that, based on the sample variance, the group Equipment Count means are not significantly different from each other.

Table 3.52: Letters Report Equipment-Price

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Equipment Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>200000</td>
<td>A</td>
<td>7</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Equipment-Price does not have a significant effect on the Equipment Count. In this case, by increasing the cost of each unit of equipment by an order of magnitude, the Total Network Cost increases by only 33%. It can be concluded that with respect to Equipment Count, that large (i.e., order of magnitude) changes in price, do not statistically significantly have an effect on the optimal topology of the network in terms of Equipment Count. So while one might wish the Equipment Count to go down with expensive Equipment Price, Equipment Count does go down but not significantly in these instances.
3.10.2.2. Pair-wise Analysis

Pair-wise analysis is next performed on pairs of factors to determine if any two factors, when combined, significantly influence the response Equipment Count. This analysis uncovers the specific combination(s) of factors that have a statistically significant effect on the response.

3.10.2.2.1 The Influence of Meters*Area on Equipment Count

From a practical perspective, the number of meters combined with area, notionally again characterizes a density. Intuitively a denser environment should drive a higher Equipment Count. The question in this section is whether there is a statistically significant difference given the densities in this set of instances under study.

To test this, for formal experimentation purposes, we state the following conjecture: all combinations of Meters and Area (denoted Meters*Area) make no difference regarding the Equipment Count. Thus, the hypothesis tested is:

**Experiment 15:**

\[ H_0 \#15: \text{The average Equipment Count} \ (Y_2) \ \text{is the same for all levels of all combinations of Meters*Area} \ (X_5). \]

\[ H_0 \#15: \text{The average Equipment Count} \ (Y_2) \ \text{is not the same for all levels of all combinations of Meters*Area} \ (X_5). \]

The Meters*Area interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.22. Note the positive slopes for Area and vertical distance between points for Meters. Particularly at the 10,000 high-end level of Meters, this suggests correlation between Area*meters (notionally density) and Equipment Count.
A two-factor analysis of variance, in Table 3.53, shows that the Meter*Area levels do not have significantly different Equipment Counts, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.2387. Thus, the null hypothesis $H_0 \ #15$ (the means are equal) is not rejected and we can assume that Meters*Area level does not have a statistically significant association with respect to Equipment Count.

Table 3.53: Two-Factor ANOVA for Meters*Area

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>$Prob &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Area</td>
<td>1</td>
<td>1</td>
<td>14.0625</td>
<td>2</td>
<td>0.2387</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Meters*Area, also characterizing levels of density, do not have a statistically significant effect on Equipment Count, for these set of instances. Perhaps for a broader range of densities, statistically significance could be shown.

3.10.2.2.2 The Influence of Meters*Lower-Bound on Equipment Count

The two-factor interaction of Meters*Lower-Bound on Equipment Count gives only an abstract intuitive notion, namely how might the Equipment Count has an effect
between two different types of equipment, differing in Lower-Bound, used in two
different levels of Meter deployments. In other words, would Equipment Count change
across combinations of Meters and Lower-Bounds. A weak intuitive expectation is
that as Lower-Bound is increased Equipment Count would decrease, and it might be
more pronounced at higher levels of Meters.

For formal experimentation purposes, the following conjecture is however statisti-
cally tested, namely, all combinations of Meters*Lower-Bound make no difference
regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 16:

$H_0$ #16: The average Equipment Count ($Y_1$) is the same for all levels of all
combinations of Meters*Lower-Bound ($X_6$).

$H_1$ #16: The average Equipment Count ($Y_1$) is not the same for all levels of all
combinations of Meters*Lower-Bound ($X_6$).

The Meters*Lower-Bound interaction has four levels, and the mean Total Network
Cost for each tested level is shown in the means plot of Figure 3.23. Note the negative
slope suggests a negative correlation of Lower-Bound on Equipment Count, and a gap
in vertical distance of the meter points suggest a correlation of Meters and Equipment
Count.

![Figure 3.23: LS Means Plot Equipment-Price vs. Equipment Count](image)

129
A two-factor analysis of variance, in Table 3.54, shows that the Meters*Lower-Bound levels do not have significantly different Equipment Counts, given a threshold of \( \alpha = 0.05 \), and a \( p \)-value equal to 0.9324. Thus, the null hypothesis \( H_0 \) (the means are equal) is not rejected and we can assume that Meters*Lower-Bound level does not have a statistically significant association with respect to Equipment Count.

Table 3.54: Two-Factor ANOVA for Meters*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>0.0625</td>
<td>0</td>
<td>0.9324</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Meters*Lower-Bound do not have a statistically significant effect on the Equipment Count.

3.10.2.2.3 The Influence of Meters*Equipment-Price on Equipment Count

For the interaction term Meters*Equipment-Price, one might expect that Equipment Price would have an effect on Equipment Count, and it might be most pronounced at increased meter levels. For formal experimentation purposes, what is tested, is that all combinations of Meters and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 17:

\[ H_0 \#17: \text{The average Equipment Count (} Y_2 \text{) is the same for all levels of all combinations of Meters*Equipment-Price (} X_7 \text{).} \]

\[ H_1 \#17: \text{The average Equipment Count (} Y_2 \text{) is not the same for all levels of all combinations of Meters*Equipment-Price (} X_7 \text{).} \]
The Meters*Equipment-Price interaction has four levels and the mean Equipment Count for each tested level is shown in the means plot of Figure 3.24. Note as Equipment Price increases, the pronounced negative slope at high levels of Meters, suggests Meters*Equipment Price negatively correlates with Equipment Count, as Meters increase.

Figure 3.24: LS Means Plot Meters*Equipment-Price Interaction vs. Equipment Count

A two-factor analysis of variance, in Table 3.55, shows that the meter*Equipment-Price levels do not have significantly different Equipment Counts, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.19. Thus, the null hypothesis $H_0 \#17$ (the means are equal) is not rejected and we can assume that Meters*Equipment-Price interaction is not statistically significant with respect to Equipment Count.

Table 3.55: Two-Factor ANOVA for Meters*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters* Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>18.0625</td>
<td>2</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters
report is not meaningful. Thus it is shown that the levels of Meters*Equipment-Price do not have a statistically significant effect on the Equipment Count.

3.10.2.2.4 The Influence of Area*Lower-Bound on Equipment Count

The two-factor interaction of Area and Lower-Bound do not give as intuitive an interpretation as does some preceding interactions. Given this, there is no intuitive expectation for this variable. One can imagine that the combination behaves as the individuals might, namely as Area increases this causes an increase in Equipment Count and as Lower-Bound increases this gives rise to lower Equipment Count, making the effect on Equipment Count indeterminate.

For formal experimentation purposes, the following conjecture is statistically tested. All combinations of Area*Lower-Bound make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 18:

\[
H_0 \ #18: \text{The average Equipment Count (} Y_2 \text{) is the same for all levels of all combinations of Area*Lower-Bound (} X_8 \text{).}
\]

\[
H_1 \ #18: \text{The average Equipment Count (} Y_2 \text{) is not the same for all levels of all combinations of Area*Lower-Bound (} X_8 \text{).}
\]

The Area*Lower-Bound interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.25. Note the negative slopes for Lower-Bound and vertical gaps for Area, suggesting some correlation.
A two-factor analysis of variance, in Table 3.56, shows that the Area*Lower-Bound levels do not have significantly different Total Network Costs, given a threshold of $\alpha=0.05$, and a $p$-value equal to 0.7998. Thus, the null hypothesis $H_0$ #18 (the means are equal) is not rejected and we can assume that Area*Lower-Bound level does not have a statistically significant association with respect to Equipment Count.

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Area*Lower-Bound do not have a statistically significant effect on the Total Network Cost.

### Table 3.56: Two-Factor ANOVA for Area*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area* Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>0.5625</td>
<td>0</td>
<td>0.7998</td>
</tr>
</tbody>
</table>

3.10.2.2.5  The Influence of Area*Equipment-Price on Equipment Count

The interaction term Area*Equipment-Price has no intuitive practical notion as does other combinations of factors. The notional expectation however is that as both the individual or combined factors are increased, so would Equipment Count.
For the formal experimentation the opposite conjecture is statistically tested. All combinations of Area and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 19:

$H_0 \ #19$: The average Equipment Count ($Y_2$) is the same for all levels of all combinations of Area*Equipment-Price ($X_9$).

$H_1 \ #19$: The average Equipment Count ($Y_2$) is not the same for all levels of all combinations of Area*Equipment-Price ($X_9$).

The Area*Equipment-Price interaction has four levels, and the mean Equipment Count for each tested level is shown in the means plot of Figure 3.26. Note the slopes and gaps, indicating some correlation.

Figure 3.26: LS Means Plot Area*Equipment-Price Interaction vs. Equipment Count

A two-factor analysis of variance, in Table 3.57, shows that the Area*Equipment-Price levels do not have significantly different Equipment Count, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.4587. Thus, the null hypothesis $H_0 \ #19$ (the means are equal) is not rejected and we can assume that Area*Equipment-Price interaction is not statistically significant with respect to Equipment Count.
Table 3.57: Two-Factor ANOVA for Area*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area*Equipment-</td>
<td>1</td>
<td>1</td>
<td>5.0625</td>
<td>1</td>
<td>0.4587</td>
</tr>
<tr>
<td>Price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Given that this interaction is not significant, the joint behavior of Area*Equipment-Price is deemed indeterminate.

3.10.2.2.6 The Influence of Lower-Bound*Equipment-Price on Equipment Count

The two-factor interaction of Lower-Bound and Equipment-Price does not give an intuitive notion when taken together. For the experiment, the conjecture is tested that all combinations of Lower-Bound*Equipment-Price make no difference regarding the Equipment Count. Thus, the hypothesis tested is:

Experiment 20:

\[ H_0 \#20: \text{The average Equipment Count } (Y_1) \text{ is the same for all levels of all combinations of Lower-Bound*Equipment-Price } (X_{10}). \]

\[ H_1 \#20: \text{The average Equipment Count } (Y_1) \text{ is not the same for all levels of all combinations of Lower-Bound*Equipment-Price } (X_{10}). \]

The Lower-Bound*Equipment-Price interaction has four levels, and the mean Equipment Count for each tested level is shown in the means plot of Figure 3.27. Note at lower Equipment Price, Lower-Bound may make some difference to Equipment Count, namely a negative correlation.
Figure 3.27: LS Means Plot Lower-Bound*Equipment-Price vs. Equipment Count

A two-factor analysis of variance, in Table 3.58, shows that the Lower-Bound*Equipment-Price levels do not have significantly different Equipment Counts, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.3718. Thus, the null hypothesis $H_0$ (the means are equal) is rejected and we can assume that Lower-Bound*Equipment-Price level does not have a statistically significant association with respect to Equipment Count.

Table 3.58: Two-Factor ANOVA for Lower-Bound*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>7.5625</td>
<td>1</td>
<td>0.3718</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Lower-Bound*Equipment-Price do not have a statistically significant effect on the Equipment Count.
3.10.2.3. Equipment Count Regression Analysis

This section presents the results of performing a least squares fit of the four main effects and six interaction terms to derive an equation for Equipment Count. The results of the regression calculations are shown below in Table 3.59. Note only Meters is significant with a $p$-value of 0.0215 ("Prob $>|t|$" in the table).

Table 3.59: Sorted Parameter Estimates Equipment Count

| Term                           | Estimate $\beta_i$ | Std Error | t Ratio | Prob $>|t|$ |
|--------------------------------|--------------------|-----------|---------|-------------|
| Meters                         | 0.000925           | 0.00028   | 3.30    | 0.0215*     |
| Area                           | 0.0065625          | 0.003505  | 1.87    | 0.1201      |
| Equipment-Price                | −1.319e−5          | 7.789e−6  | −1.69   | 0.1510      |
| (Meters-7500)* (Equipment-Price-110000) | −4.722e−9    | 3.116e−9  | −1.52   | 0.1900      |
| (Meters-7500)* (Area-300)      | 1.875e−6           | 1.402e−6  | 1.34    | 0.2387      |
| Lower-Bound                    | −0.007639          | 0.007789  | −0.98   | 0.3718      |
| (Lower-Bound-110)* (Equipment-Price-110000) | 8.4877e−8     | 8.654e−8  | 0.98    | 0.3718      |
| (Area-300)* (Equipment-Price-110000) | −3.125e−8     | 3.894e−8  | −0.80   | 0.4587      |
| (Area-300)*(Lower-Bound-110)   | 1.0417e−5         | 0.000039  | 0.27    | 0.7998      |
| (Meters-7500)* (Lower-Bound-110) | −2.778e−7     | 3.116e−6  | −0.09   | 0.9324      |
Table 3.60: Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>188.62500</td>
<td>18.8625</td>
<td>2.3990</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>39.31250</td>
<td>7.8625</td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>15</td>
<td>227.93750</td>
<td>0.1731</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.61: Summary of Fit

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

The analysis of variance shows a poor fit of the data to the regression line, with a $p$-value ($Prob > F$) of 0.1731. While the RSquare value of 0.8275 tends to indicate a good fit. However, the adjusted RSquare value, which allows for the large number of parameters in the model, is computed to be 0.482, thereby confirming the poor fit of the model.
The equation plotted against the actual data is shown in Figure 3.28. The 95% confidence dotted lines do not closely straddle the regression line, indicating a poor confidence of fit. Also, these curves straddle the (horizontal dotted) mean line, thereby further indicating the observed data may just be reflecting random data about the mean.

The fitted equation for Equipment Count is shown in Figure 3.29. From table 3.59, the first term, Meters, is the only term with statistical significance.
Equipment Count =
\[
1.82291666666667 \\
+ 0.000925 \times \text{Meters} \\
+ 0.0065625 \times \text{Area} \\
+ -0.00763888888889 \times \text{Lower Bound} \\
+ -0.0000131944444 \times \text{Equipment Price} \\
\left( \frac{\text{Meters} - 7500}{7500} \right) \\
\times \left( \frac{\text{Area} - 300}{300} \times 0.000001875 \right) \\
\left( \frac{\text{Meters} - 7500}{7500} \right) \\
\times \left( \frac{\text{Lower Bound} - 110}{110} \times -2.77777777778 \times 10^{-7} \right) \\
\left( \frac{\text{Meters} - 7500}{7500} \right) \\
\times \left( \frac{\text{Equipment Price} - 110000}{110000} \times -4.7222222222 \times 10^{-9} \right) \\
\left( \frac{\text{Area} - 300}{300} \right) \\
\times \left( \frac{\text{Lower Bound} - 110}{110} \times 0.00001041666667 \right) \\
\left( \frac{\text{Area} - 300}{300} \right) \\
\times \left( \frac{\text{Equipment Price} - 110000}{110000} \times 0.0000003125 \right) \\
\left( \frac{\text{Lower Bound} - 110}{110} \right) \\
\times \left( \frac{\text{Equipment Price} - 110000}{110000} \times 8.48765432099 \times 10^{-8} \right)
\]
3.10.2.4. Summary for Equipment Count

The summary of experimental hypothesis testing is shown in Table 3.62. Here all tests, both interactions and main effects are not statistically significant in modeling Equipment Count.

Table 3.62: Summary Results for Equipment Count

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ($H_0$)</th>
<th>$Prob &gt; t$</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Equipment Counts are equal for Area</td>
<td>0.1201</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>13</td>
<td>Equipment Counts are equal for L.Bound</td>
<td>0.3718</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>14</td>
<td>Equipment Counts are equal for E.Price</td>
<td>0.1510</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>15</td>
<td>Equipment Counts are equal for Meters* Area</td>
<td>0.2387</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>16</td>
<td>Equipment Counts are equal for Meters* L.Bound</td>
<td>0.9324</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>17</td>
<td>Equipment Counts are equal for Meters* E.Price</td>
<td>0.1900</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>18</td>
<td>Equipment Counts are equal for Area* L.Bound</td>
<td>0.7998</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>19</td>
<td>Equipment Counts are equal for Area* E.Price</td>
<td>0.4587</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>20</td>
<td>Equipment Counts are equal for L.Bound* E.Price</td>
<td>0.3718</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>
For this set of instances, there is insufficient evidence that any tested factor demonstrably has an effect on Equipment Count (Area, Meters, Equipment-Price, Lower-Bound, nor any of their interactions).

Practitioners may have intuitive notions that the factors tested here can significantly has an effect on Equipment Count, but the results here, for this data set, is not supportive of that conjecture.

3.10.3. Average Link Distance

When designing a communications network, the reliability of communications links is important. Reliability of links is a function of many aspects, but one typically dominant aspect is the strength of the communications signal. A strong signal is typically detected with a much higher reliability than a weak signal. As communications links become longer, signal strength typically becomes weaker. Longer links are generally less reliable than shorter links.

A figure of merit of a design is the average length of communications links. The outcome of a design’s average link distance may be influenced by initial conditions and designer decisions at the start of a design. The following deals with the association of the design input factors on the response Average Link Distance.

3.10.3.1. Single-Factor Analysis

The following section performs single-factor ANOVA on experimental observations to explore the effect, if any, of each of four factors on the response Average Link Distance. The four factors are number of Meters, Area covered, Lower-Bound on link traffic, and Equipment-Price. The Average Link Distance in an important network design characteristic. As a broad generalization, shorter links are desirable over longer links because they tend to have higher reliability. Therefore distance is reflected in the model’s formulation as a “cost” on the arcs of the model to drive down link distances. The distances are computed as the Euclidean distance, in grid points, between all
equipment (meters, repeaters and collectors).

3.10.3.1.1 The Influence of Meters on Average Link Distance

From a practical perspective, the number of Meters served would intuitively increase density across the Area under test, and thus tend to reduce the Average Link Distance. The conjectural expectation therefore is that, as the level of Meters is changed (increased), the Average Link Distance changes (decreases).

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Meters has no relationship with Average Link Distance. Thus, the hypothesis tested is:

Experiment 21:

\( H_0 \ #21: \) The Average Link Distance \( (Y_3) \) is the same for both levels of Meters \( (X_1) \).

\( H_1 \ #21: \) The Average Link Distance \( (Y_3) \) is not the same for both levels of Meters \( (X_1) \).

In this study, the number of Meters has two levels, and the mean Average Link Distance for each tested level is shown in the least square means plot of Figure 3.30. The negative slope suggest negative correlation of Meters on Average Link Distance.
A one-factor analysis of variance, in Table 3.63, shows that the Meter levels have significantly different Average Link Distances, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.0001. Thus, the null hypothesis $H_0 \neq 21$ (the means are equal) is rejected and we can assume that Meters level has a statistically significant association with respect to Average Link Distance.

Table 3.63: ANOVA for Meters

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>$Prob &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters</td>
<td>1</td>
<td>1</td>
<td>470.4561</td>
<td>120</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

As further evidence shown in Table 3.64, the least-squares-means Student’s $t$-test compares the means at the 10000 versus 5000 meter levels, and gives the result, for a two-tailed test, of $Prob > |t| = 0.0001$. In the graphic, the sampling Student’s $t$ distribution curve is shown relative to the $t$-ratio marker indicating the computed $t$-ratio. This illustrates how extremely unequal the observed means are, thus leading to a low $p$-value (the probability that the two observed means are equal), and conclusion the means are not equal.
Table 3.64: Student $t$-test Comparing 10000 with 5000 Meters for Average Link Distance

<table>
<thead>
<tr>
<th>Difference</th>
<th>$t$ Ratio</th>
<th>Std Err Dif</th>
<th>DF</th>
<th>Upper CL Dif</th>
<th>Lower CL Dif</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-10.845$</td>
<td>$-10.9767$</td>
<td>$0.988$</td>
<td>$5$</td>
<td>$-8.305$</td>
<td>$-13.385$</td>
<td>$0.95$</td>
</tr>
</tbody>
</table>

The letters report, shown in Table 3.65, assigns the observations for each level of Meters into statistically separate groups indicating that, based on the sample variance, the group Equipment Count means are significantly different from each other.

Table 3.65: Letters Report Meters

<table>
<thead>
<tr>
<th>Level</th>
<th>$t$-test Group</th>
<th>Average Link Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>A</td>
<td>30.0</td>
</tr>
<tr>
<td>10000</td>
<td>B</td>
<td>19.2</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Meters has a statistically significant effect on the Average Link Distance. As the number of Meters increase, there is an increase in infrastructure needed to connect and transport the increased traffic. In this case, doubling the number of Meters, amounts to a $(30 - 19.2/19.2 =) 57\%$ change (decrease) in Average Link Distance across changes in the other factors of Area, Lower-Bound and Equipment-Price.

In practical networks, as the number of meters increase, the Average Link Distance intuitively decreases. The experimentation for this set of data, bears this out.
3.10.3.1.2 The Influence of Area on Average Link Distance

The area covered in a deployment is an important factor in terms of the type of equipment needed for the network. In practical networks, the need to span large areas requires long range equipment, or more hops. In this experiment, the expectation is that as the level of Area is increased, the Average Link Distance increases (since we have not incorporated a range limitation on link distance that might drive a long hop into multiple hops).

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Area makes no difference regarding the Average Link Distance. Thus, the hypothesis tested is:

Experiment 22:

$H_0$ #22: The Average Link Distance ($Y_3$) is the same for both levels of Area ($X_2$).

$H_1$ #22: The Average Link Distance ($Y_3$) is not the same for both levels of Area ($X_2$).

The Area has two levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.31. Note the strong positive slope suggesting Area has a pronounced effect on Average Link Distance.
A one-factor analysis of variance shown in Table 3.66, confirms that the Area levels do have significantly different Average Link Distances given a threshold of $\alpha = 0.05$, and a $p$-value equal to $< 0.0001$. Thus, the null hypothesis $H_0 \#22$ (the means are equal) is rejected and we can assume that Area level has a statistically significant association with respect to Average Link Distance.

Table 3.66: ANOVA for Area

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>$Prob &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>1</td>
<td>1</td>
<td>6120.7152</td>
<td>1,568</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

As further evidence shown in Table 3.67, the least-squares-means Student’s $t$-test compares the means at the 500 versus 100 Area levels, and gives the result, for a two-tailed test, of $Prob > |t| \leq 0.0001$. The sampling student’s $t$ distribution curve is shown relative to the line indicating the computed $t$-ratio. The asterisk indicates that this value is statistically significant (less than the $\alpha = 0.05$ threshold). In the table’s graphic, the sampling Student’s $t$ distribution curve is shown, along with the $t$-ratio marker on the abscissa showing the relative location of the computed $t$-ratio.
The distance of the computed $t$-ratio marker from the mean of the $t$-distribution curve illustrates how extremely unequal the observed means are, thus leading to a low $p$-value (the probability that the two observed means are equal).

Table 3.67: Student’s $t$-test for Area Comparing 500 with 100 Areas

<table>
<thead>
<tr>
<th>Difference</th>
<th>$t$ Ratio</th>
<th>Std Err Dif</th>
<th>DF</th>
<th>Upper CL Dif</th>
<th>Lower CL Dif</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.1175</td>
<td>39.59244</td>
<td>0.9880</td>
<td>5</td>
<td>41.6572</td>
<td>36.5778</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**Prob &gt;</td>
<td>**</td>
<td><strong>Prob &gt;</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$</td>
<td>t</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The letters report, shown in Table 3.68, assigns the observations for each level of Area into two groups confirming that, based on the sample variance, the group Total Network Cost means are significantly different from each other.

Table 3.68: Letters Report Area

<table>
<thead>
<tr>
<th>Level</th>
<th>$t$-test Group</th>
<th>Average Link Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>A</td>
<td>44.2</td>
</tr>
<tr>
<td>100</td>
<td>B</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Area does have a statistically significant effect on the Average Link Distance. As the Area increases, there this is a 774% increase in Average Link Distance. Thus, with a 5-fold increase in Area, this set of data shows a statistically significant 7-fold increase in Average Link Distance.

3.10.3.1.3 The Influence of Lower-Bound on Average Link Distance

Intuitively, the more stringent (higher) the Lower-Bound, the more links (and sites) can be eliminated from a network, thereby increasing potentially the Average Link...
Distance needed for links to transport traffic. Consequently, in this experiment, the notional expectation is that, as the level of Lower-Bound is changed, the Average Link Distance changes.

For formal experimentation purposes, the opposite conjecture is statistically tested: the level of Lower-Bound makes no difference regarding the Average Link Distance. Thus, the hypothesis tested is:

Experiment 23:

$H_0$ #23: The Average Link Distance ($Y_3$) is the same for both levels of Lower-Bound ($X_3$).

$H_1$ #23: The Average Link Distance ($Y_3$) is not the same for both levels of Lower-Bound ($X_3$).

The Lower-Bound has two levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.32. The flat line suggests no correlation.

![Figure 3.32: LS Means Plot Lower-Bound vs. Average Link Distance](image)

A one-factor analysis of variance, in Table 3.69, shows that the Lower-Bound
levels do not have significantly different Average Link Distance, given a threshold of $\alpha = 0.05$, and a $p$-value equal to $< 0.7827$. Thus, the null hypothesis $H_0 \#23$ (the means are equal) is not rejected and we can assume that Lower-Bound level does not have a statistically significant association with respect to Total Network Cost.

Table 3.69: ANOVA for Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>0.3306</td>
<td>0</td>
<td>0.7827</td>
</tr>
</tbody>
</table>

Shown in Table 3.70, the least-squares-means Student’s $t$-test compares the means at the 20 versus 200 Lower-Bound levels, and gives the result, for a two-tailed test, of $\text{Prob} > |t| = 0.7827$. In this graphic, the sampling Student’s $t$ distribution curve illustrates the area representing how likely the probability is that the observed means are equal, thus leading to a high $p$-value, and insufficient evidence to reject the null hypothesis.

Table 3.70: Student’s $t$-test for Lower-Bound Comparing 200 with 20

<table>
<thead>
<tr>
<th>Difference</th>
<th>0.2875</th>
<th>$t$ Ratio</th>
<th>0.290991</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>0.9880</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>2.8272</td>
<td>$\text{Prob} &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>−2.2522</td>
<td>$\text{Prob} &gt; t$</td>
<td>0.3914</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>$\text{Prob} &lt; t$</td>
<td>0.6086</td>
</tr>
</tbody>
</table>

The $t$-test letters report assigns the observations for each level of Meters into statistically separate groups, shown in Table 3.71, indicating that, based on the sample variance, the group Average Link Distance means are not significantly different from each other.
Thus it is shown that the level of Lower-Bound does not have a statistically significant effect on the Average Link Distance. As the Lower-Bound increases, there is not a statistically significant increase in the Average Link Distances traveled. In this case, increasing the Lower-Bound by an order of magnitude accounts for only an increase of 1% in Average Link Distance. So a counterintuitive result is obtained from this experiment, namely increasing the Lower-Bound does not significantly change Average Link Distance.

3.10.3.1.4 The Influence of Equipment-Price on Average Link Distance

From a practical perspective, increasing Equipment-Price should has an effect on Average Link Distance by forcing longer links versus more equipment when feasible, given there is a small cost penalty for distance in the model but a larger cost penalty for Equipment-Price. Thus, it is expected that as the Equipment-Price increases, the Average Link Distance is reduced. In this section we quantify this notion by testing an increase of Equipment-Price ten-fold.

For formal experimentation purposes, the following conjecture is statistically tested: the level of Equipment-Price makes no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Link Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>A</td>
<td>24.8</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>24.5</td>
</tr>
</tbody>
</table>
Experiment 24:

$H_0$ #24: Average Link Distance ($Y_3$) is the same for both levels of Equipment-Price ($X_4$).

$H_1$ #24: Average Link Distance ($Y_3$) is not the same for both levels of Equipment-Price ($X_4$).

The Equipment-Price has two levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.33. The slight slope indicates a possible correlation.

![Figure 3.33: LS Means Plot Lower-Bound vs. Average Link Distance](image)

A one-factor analysis of variance, in Table 3.72, shows that the Equipment-Price levels do have significantly different Average Link Distances, given a threshold of $\alpha = 0.05$, and a $p$-value equal to $> 0.0145$. Thus, the null hypothesis $H_0$ #24 (the means are equal) is rejected and we can assume that Equipment-Price level has a statistically significant association with respect to Average Link Distance.
Table 3.72: ANOVA for Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>52.5625</td>
<td>13</td>
<td>0.0145</td>
</tr>
</tbody>
</table>

The least-squares-means Student’s $t$-test, shown in Table 3.73, compares the means at the 2,000 versus 200,000 levels, and gives the result, for a two-tailed test, of $\text{Prob} \geq |t| \geq 0.0145$. The graphic illustrates the critical regions representing the probability of a Type I error, and likelihood of concluding incorrectly that the means are equal. In the table’s Student’s $t$ test graphic, it is seen how far to the right the $t$-ratio marker is, indicating how extremely unequal the observed means are, thus leading to a low $p$-value, and rejection of the null hypothesis.

Table 3.73: Student’s $t$-test for Equipment-Price Comparing 200,000 with 20,000 Prices

<table>
<thead>
<tr>
<th>Difference</th>
<th>3.62500</th>
<th>$t$ Ratio</th>
<th>3.669012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Err Dif</td>
<td>0.98800</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>6.16475</td>
<td>$\text{Prob} \geq</td>
<td>t</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>1.08525</td>
<td>$\text{Prob} &gt; t$</td>
<td>0.0072$^*$</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
<td>$\text{Prob} &lt; t$</td>
<td>0.9928</td>
</tr>
</tbody>
</table>

Lastly, a $t$-test letters report, shown in Table 3.74, indicates that, based on the sample variance, the group Average Link Distance means are significantly different from each other.
Table 3.74: Letters Report Equipment-Price

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Link Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>200000</td>
<td>A</td>
<td>26.4</td>
</tr>
<tr>
<td>20000</td>
<td>B</td>
<td>22.8</td>
</tr>
</tbody>
</table>

Thus it is shown that the level of Equipment-Price does have a significant effect on the Average Link Distance. In this case, by increasing the cost of each unit of equipment by an order of magnitude, the Average Link Distance increases by 16%. It can be concluded that with respect to Average Link Distance, that large (i.e., order of magnitude) changes in price, do statistically significantly effect the Average Link Distance.

3.10.3.2. Pair-wise Analysis

Pair-wise analysis is next performed on pairs of factors to determine if any two factors, when combined, significantly influence the response Average Link Distance. This analysis uncovers the specific combination(s) of factors that have a statistically significant effect on the response.

3.10.3.2.1 The Influence of Meters*Area on Average Link Distance

From a practical perspective, the number of meters combined with area, notionally characterizes a density. Intuitively a denser environment of Meters should drive a lower Average Link Distance. The question in this section is whether there is a statistically significant difference given the densities in this set of instances.

To test this, for formal experimentation purposes, we state the following conjecture: all combinations of Meters and Area (denoted Meters*Area) make no difference regarding the Average Link Distance in the optimally designed network. Thus, the hypothesis tested is:
Experiment 25:

$H_0$ #25: The Average Link Distance ($Y_3$) is the same for all levels of all combinations of Meters*Area ($X_5$).

$H_1$ #25: The Average Link Distance ($Y_3$) is not the same for all levels of all combinations of Meters*Area ($X_5$).

The Meters*Area interaction has four levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.34. Note the strong positive slope for variation in Area, and the vertical distance gap for variation in Meters, both indicating correlation for this interaction.

Figure 3.34: LS Means Plot Meters*Area Interaction vs. Average Link Distance

A two-factor analysis of variance, in Table 3.75, shows that the Meter*Area levels do have significantly different Average Link Distances, given a threshold of $\alpha=0.05$, and a $p$-value equal to 0.0004. Thus, the null hypothesis $H_0$ #25 (the means are equal) is not rejected and we can assume that Meters*Area level has a statistically significant association with respect to Average Link Distance.
Table 3.75: Two-Factor ANOVA for Meters*Area

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Area</td>
<td>1</td>
<td>1</td>
<td>273.737</td>
<td>70</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report, is indicated. As shown in Table 3.76, based on the sample variance, the group Average Link Distance means fall into three significantly different groups.

Table 3.76: Letters Report - Meters*Area for Average Link Distance

<table>
<thead>
<tr>
<th>Level</th>
<th>t-test Group</th>
<th>Average Link Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000,500</td>
<td>A</td>
<td>53.7</td>
</tr>
<tr>
<td>10,000,500</td>
<td>B</td>
<td>34.6</td>
</tr>
<tr>
<td>5,000,100</td>
<td>C</td>
<td>6.3</td>
</tr>
<tr>
<td>10,000,100</td>
<td>C</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Thus it is shown that the levels of Meters*Area, also representing levels of density, have a statistically significant effect on the Average Link Distance. As the number of Meters*Area increase, Average Link Distance is driven in major increments by Area, and minor increments within Area, by Meters. Each increment from lowest to highest increases Average Link Distance incrementally from the lowest by 68%, then as Area is incremented, by 446%, and finally with the last increment of Meters, by 55%. Because this interaction is significant, this gives support to the fact that the two individual factors are also significant (as shown in the prior sections).

These results are interesting in that there is a statistically significant difference at three of the four density levels, the largest increase coming as Area is increased five-fold, then within that at the highest level of Area, as Meters are doubled. The practical implication is that both factors play a significant role in the Average Link
Distance, individually and jointly, with Area being the major driver for these sets of tests.

3.10.3.2.2 The Influence of Meters*Lower-Bound on Average Link Distance

The two-factor interaction of Meters and Lower-Bound (denoted Meters*Lower-Bound) gives only an abstract intuitive notion of four different types of lower-bound-constraining pairs of equipment used for a link. There is no intuitive expectation for this combination of factors.

For formal experimentation purposes, the following conjecture is however statistically tested: all combinations of Meters*Lower-Bound make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 26:

\[ H_0 \ #26: \text{The Average Link Distance (} Y_3 \text{) is the same for all levels of all combinations of Meters*Lower-Bound (} X_6 \text{).} \]

\[ H_1 \ #26: \text{The Average Link Distance (} Y_3 \text{) is not the same for all levels of all combinations of Meters*Lower-Bound (} X_6 \text{).} \]

The Meters*Lower-Bound interaction has four levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.35. With a strong slope for Area and large gap, particularly on the high end of Area, for Meters, this suggests correlation of this interaction to Average Link Distance.
However, a two-factor analysis of variance, in Table 3.77, shows that the Meters*Lower-Bound levels do not have significantly different Average Link Distances, given a threshold of $\alpha=0.05$, and a $p$-value equal to 0.222. Thus, the null hypothesis $H_0$ (the means are equal) is not rejected and we can assume that Meters*Lower-Bound level does not have a statistically significant association with respect to Average Link Distance.

Table 3.77: Two-Factor ANOVA for Meters*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>7.59</td>
<td>2</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Meters*Lower-Bound do not have a statistically significant effect on the Average Link Distance.

3.10.3.2.3 The Influence of Meters*Equipment-Price on Average Link Distance

For the interaction term Meters*Equipment-Price, although there is no practical notion for the combination, it would seem that with both factors increasing individually,
one would counter-balance the other, with the net effect on Average Link Distance to be indeterminable.

For formal experimentation purposes, what is tested, is that all combinations of Meters and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 27:

\[ H_0 \#27: \text{The Average Link Distance (} Y_3 \text{) is the same for all levels of all combinations of Meters*Equipment-Price (} X_7 \text{).} \]

\[ H_1 \#27: \text{The Average Link Distance (} Y_3 \text{) is not the same for all levels of all combinations of Meters*Equipment-Price (} X_7 \text{).} \]

The Meters*Equipment-Price interaction has four levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.36. Slight slope and vertical gap suggests correlation.

![Figure 3.36: LS Means Plot Meters*Equipment-Price Interaction vs. Average Link Distance](image)

However, a two-factor analysis of variance, in Table 3.78, shows that the Meter*Equipment-Price levels do not have significantly different mean Average Link Distances, given a threshold of \( \alpha = 0.05 \), and a \( p \)-value equal to 0.1822. Thus, the null hypothesis \( H_0 \#27 \)
(the means are equal) is not rejected and we can assume that Meters*Equipment-Price interaction is not statistically significant with respect to Average Link Distance.

Table 3.78: Two-Factor ANOVA for Meters*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meters*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>9.3636</td>
<td>2</td>
<td>0.1822</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Meters*Equipment-Price do not have a statistically significant effect on the Average Link Distance.

3.10.3.2.4 The Influence of Area*Lower-Bound on Average Link Distance

The two-factor interaction of Area and Lower-Bound (denoted Area*Lower-Bound) do not give an intuitive feeling for a likely result. Given this, there is no intuitive expectation for this interaction. One can imagine that the combination behaves as the individual counterparts do. With Area increasing, causing increased Average Link Distance and Lower-Bound increasing giving rise to increased Average Link Distance, when both are increasing, Average Link Distance may be thought to increase.

For formal experimentation purposes, the following conjecture is statistically tested. All combinations of Area*Lower-Bound make no difference regarding the mean Total Network Cost. Thus, the hypothesis tested is:

Experiment 28:

\[ H_0 \#28: \text{The Average Link Distance (}\hat{Y}_3\text{) is the same for all levels of all combinations of Area*Lower-Bound (}\hat{X}_8\text{).} \]

\[ H_1 \#28: \text{The Average Link Distance (}\hat{Y}_3\text{) is not the same for all levels of all combinations of Area*Lower-Bound (}\hat{X}_8\text{).} \]
The Area*Lower-Bound interaction has four levels, and the mean Total Network Cost for each tested level is shown in the means plot of Figure 3.37. Note the large vertical gap in Area, and slightly counter posing slopes for Lower-Bound.

Figure 3.37: LS Means Plot Area*Lower-Bound vs. Average Link Distance

A two-factor analysis of variance, in Table 3.79, shows that the Area*Lower-Bound levels do not have significantly different mean Average Link Distances, given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.589. Thus, the null hypothesis $H_0 \#28$ (the means are equal) is not rejected and we can assume that Area*Lower-Bound level is not statistically significant with respect to Average Link Distance.

Table 3.79: Two-Factor ANOVA for Area*Lower-Bound

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area*Lower-Bound</td>
<td>1</td>
<td>1</td>
<td>1.2996</td>
<td>0</td>
<td>0.589</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Area*Lower-Bound do not have a statistically significant effect on the Total Network Cost.
3.10.3.2.5 The Influence of Area*Equipment-Price on Average Link Distance

The interaction term Area*Equipment-Price has no intuitive practical notion as does other combinations of main effect factors. However, the expectation is that as both the individual factors are increased, so would Average Link Distance.

For the formal experimentation the opposite conjecture is statistically tested. All combinations of Area and Equipment-Price make no difference regarding the Total Network Cost. Thus, the hypothesis tested is:

Experiment 29:

\[ H_0 \#29: \text{The Average Link Distance (} Y_3 \text{) is the same for all levels of all combinations of Area*Equipment-Price (} X_9 \text{).} \]

\[ H_0 \#29: \text{The Average Link Distance (} Y_3 \text{) is not the same for all levels of all combinations of Area*Equipment-Price (} X_9 \text{).} \]

The Area*Equipment-Price interaction has four levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.38. Large gap and positive slopes portend correlation.

![Figure 3.38: LS Means Plot Area*Equipment-Price Interaction vs. Average Link Distance](image)

A two-factor analysis of variance, in Table 3.80, shows that the Area*Equipment-
Price levels do not have significantly different mean Average Link Distances (although close), given a threshold of $\alpha = 0.05$, and a $p$-value equal to 0.0772. Thus, the null hypothesis $H_0$ (#29 (the means are equal) is rejected and we can assume that Area*Equipment-Price interaction is not statistically significant with respect to Average Link Distance.

Table 3.80: Two-Factor ANOVA for Area*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area*Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>19.2282</td>
<td>5</td>
<td>0.0772</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Given that this interaction is not significant, the joint behavior of Area*Equipment-Price is deemed indeterminate.

3.10.3.2.6 The Influence of Lower-Bound*Equipment-Price on Average Link Distance

The two-factor interaction of Lower-Bound and Equipment-Price does not give an intuitive notion when taken together. Given this, there is no intuitive expectation for the combination, thus Average Link Distance is indeterminate, when both Lower-Bound and Equipment-Price change.

For the experiment, the conjecture tested is that all combinations of Lower-Bound*Equipment-Price make no difference regarding Average Link Distance. Thus, the hypothesis tested is:

Experiment 30:
$H_0$ #30: The Average Link Distance ($Y_1$) is the same for all levels of all combinations of Lower-Bound*Equipment-Price ($X_{10}$).

$H_1$ #30: The Average Link Distance ($Y_1$) is not the same for all levels of all combinations of Lower-Bound*Equipment-Price ($X_{10}$).

The Lower-Bound*Equipment-Price interaction has four levels, and the mean Average Link Distance for each tested level is shown in the means plot of Figure 3.39.

![Figure 3.39: LS Means Plot Lower-Bound*Equipment-Price vs. Average Link Distance](image)

A two-factor analysis of variance, in Table 3.81, shows that the Lower-Bound*Equipment-Price levels do not have significantly different mean Average Link Distances, given a threshold of $\alpha= 0.05$, and a $p$-value equal to 0.7864. Thus, the null hypothesis $H_0$ #30 (the means are equal) is not rejected and we can assume that Lower-Bound*Equipment-Price level does not have a statistically significant association with respect to Average Link Distance.
Table 3.81: Two-Factor ANOVA for Lower-Bound*Equipment-Price

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Bound* Equipment-Price</td>
<td>1</td>
<td>1</td>
<td>0.3192</td>
<td>0</td>
<td>0.7864</td>
</tr>
</tbody>
</table>

Given this result, a post-hoc comparison of means using Tukey HDS with a letters report is not meaningful. Thus it is shown that the levels of Lower-Bound*Equipment-Price do not have a statistically significant effect on the Average Link Distance.

3.10.3.3. Average Link Distance Regression Analysis

This section presents the results of performing a least squares fit of the four main effects and six interaction terms to derive an equation for Average Link Distance. The results of the regression calculations are shown below in Table 3.84.

Table 3.82: Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>6,955.6022</td>
<td>695.560</td>
<td>178.1382</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>19.5231</td>
<td>3.905</td>
<td></td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>15</td>
<td>6,975.1252</td>
<td></td>
<td></td>
<td>&lt; .0001*</td>
</tr>
</tbody>
</table>
Table 3.84: Sorted Parameter Estimates Average Link Distance

| Term                          | Estimate $\beta_i$ | Std Error | $t$-Ratio | $Prob > |t|$ |
|-------------------------------|--------------------|-----------|-----------|----------|
| Area                          | 0.0977938          | 0.00247   | 39.59     | <.0001*  |
| Meters                        | -0.002169          | 0.000198  | -10.98    | 0.0001*  |
| (Meters-7500)*                | -8.272e-6          | 9.88e-7   | -8.37     | 0.0004*  |
| (Area-300)                    |                    |           |           |          |
| Equipment-Price               | 2.0139e-5          | 5.489e-6  | 3.67      | 0.0145*  |
| (Area-300)*                   |                    |           |           |          |
| (Equipment-Price-110000)      | 6.0903e-8          | 2.744e-8  | 2.22      | 0.0772   |
| (Meters-7500)*                | -3.4e-9            | 2.196e-9  | -1.55     | 0.1822   |
| (Equipment-Price-110000)      |                    |           |           |          |
| (Meters-7500)*                | 3.0611e-6          | 2.196e-6  | 1.39      | 0.2220   |
| (Lower-Bound-110)             |                    |           |           |          |
| (Area-300)*                   | -1.583e-5          | 2.744e-5  | -0.58     | 0.5890   |
| (Lower-Bound-110)             |                    |           |           |          |
| (Lower-Bound)                 | 0.0015972          | 0.005489  | 0.29      | 0.7827   |
| (Lower-Bound-110)*            | 1.7438e-8          | 6.099e-8  | 0.29      | 0.7864   |
The analysis of variance shows a good fit of the regression line to the data, with a $p$-value ($Prob > F$) of < .0001 from Table 3.82 and the RSquare value of 0.997 from Table 3.83. The adjusted RSquare value computed is 0.992, further confirms the goodness of fit.
The equation plotted against actual data is shown in Figure 3.40. The 95% confidence dotted lines do closely straddle the regression line, indicating a good confidence of fit. Also, the confidence curves do not straddle the mean line, thereby further indicating the derived equation is a good model for the observed data. The fitted equation for Average Link Distance is shown in Figure 3.41.
Average Link Distance =

\[
9.14840277777778 + -0.002169 \times \text{Meters} + 0.09779375 \times \text{Area} + 0.00159722222222 \times \text{Lower Bound} + 0.00002013888889 \times \text{Equipment Price} \\
\left( \text{Meters} - 7500 \right) + \left( \text{Area} - 300 \right) + -0.0000082725 \\
\left( \text{Meters} - 7500 \right) + \left( \text{Lower Bound} - 110 \right) \times 0.00000306111111 \\
\left( \text{Meters} - 7500 \right) + \left( \text{Equipment Price} - 110000 \right) + -0.0000000034 \\
\left( \text{Area} - 300 \right) + \left( \text{Lower Bound} - 110 \right) + -0.00001583333333 \\
\left( \text{Area} - 300 \right) + \left( \text{Equipment Price} - 110000 \right) + 6.09027777778 \times 10^{-8} \\
\left( \text{Lower Bound} - 110 \right) + \left( \text{Equipment Price} - 110000 \right) + 1.74382716049 \times 10^{-8}
\]

Figure 3.41: Prediction Model for Average Link Distance
3.10.3.4. Summary for Average Link Distance

The summary of experimental hypothesis testing is shown in Table 3.85. When viewing all tests holistically, firstly consider the interactions, then the main effects. Therefore, the only important interaction is Meters*Area; other interactions are not statistically significant. The important main effects are Meters, Area and Equipment-Price. Note the important interaction of Meters*Area supports and strengthens the conclusion of the importance of the main effects at least for Area and Meters. Thus, for this set of solution instances, Meters, Area and Equipment-Price are significant factors in the resulting Average Link Distance.

Table 3.85: Summary Results for Equipment Count

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis (H₀)</th>
<th>Prob &gt; t</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Average Link Distances are equal for Meters</td>
<td>0.0001</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>22</td>
<td>Average Link Distances are equal for Area</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>23</td>
<td>Average Link Distances are equal for L.Bound</td>
<td>0.7827</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>24</td>
<td>Average Link Distances are equal for E.Price</td>
<td>0.0145</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>25</td>
<td>Average Link Distances are equal for Meters*Area</td>
<td>0.0004</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>26</td>
<td>Average Link Distances are equal for Meters*L.Bound</td>
<td>0.2220</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>27</td>
<td>Average Link Distances are equal for Meters*E.Price</td>
<td>0.1822</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>

*Continued on next page*
Practitioners can be reassured, by these results, of their intuitive notion that Meters and Area do have an impact on Average Link Distances. It is instructive that, counterintuitively, a Lower-Bound does not statistically significantly have an effect on Average Link Distance. Thus, at least for this set of instances, setting a Lower-Bound is not statistically significant in driving Average Link Distance.

### 3.10.4. Summary Statistical Analysis

Table 3.86 lists all hypothesis tests performed, and their results. The table lists the hypothesis tested, the \( p \)-value (labeled “\( \text{Prob} > t \)”), the conclusion of the test (whether the null hypothesis is rejected or not), and whether the factor or interaction is statistically significant or not.

Table 3.86: Summary of Experimental Observations and Hypothesis Testing Results

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ( (H_0) )</th>
<th>( \text{Prob} &gt; t )</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total Network Costs are equal for Meters</td>
<td>0.0002</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>28</td>
<td>Average Link Distances are equal for Area* L.Bound</td>
<td>0.5890</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>29</td>
<td>Average Link Distances are equal for Area* E.Price</td>
<td>0.0772</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>30</td>
<td>Average Link Distances are equal for L.Bound* E.Price</td>
<td>0.7864</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>

Continued on next page
Table 3.86 – Continued from previous page

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ((H_0))</th>
<th>(Prob &gt; t)</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Total Network Costs are equal for Area</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>3</td>
<td>Total Network Costs are equal for L.Bound</td>
<td>0.4335</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>4</td>
<td>Total Network Costs are equal for E.Price</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>5</td>
<td>Total Network Costs are equal for Meters* Area</td>
<td>0.0060</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>6</td>
<td>Total Network Costs are equal for Meters* L.Bound</td>
<td>0.2244</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>7</td>
<td>Total Network Costs are equal for Meters* E.Price</td>
<td>0.0227</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>8</td>
<td>Total Network Costs are equal for Area* L.Bound</td>
<td>0.4175</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>9</td>
<td>Total Network Costs are equal for Area* E.Price</td>
<td>0.0005</td>
<td>Reject</td>
<td>Effect is statistically significant</td>
</tr>
<tr>
<td>10</td>
<td>Total Network Costs are equal for L.Bound* E.Price</td>
<td>0.6111</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ((H_0))</th>
<th>(\text{Prob} &gt; t)</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Equipment Counts are equal for</td>
<td>Meters</td>
<td>0.0215</td>
<td>Reject</td>
</tr>
<tr>
<td>12</td>
<td>Equipment Counts are equal for</td>
<td>Area</td>
<td>0.1201</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>13</td>
<td>Equipment Counts are equal for</td>
<td>L.Bound</td>
<td>0.3718</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>14</td>
<td>Equipment Counts are equal for</td>
<td>E.Price</td>
<td>0.1510</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>15</td>
<td>Equipment Counts are equal for</td>
<td>Meters* Area</td>
<td>0.2387</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>16</td>
<td>Equipment Counts are equal for</td>
<td>Meters* L.Bound</td>
<td>0.9324</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>17</td>
<td>Equipment Counts are equal for</td>
<td>Meters* E.Price</td>
<td>0.1900</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>18</td>
<td>Equipment Counts are equal for</td>
<td>Area* L.Bound</td>
<td>0.7998</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>19</td>
<td>Equipment Counts are equal for</td>
<td>Area* E.Price</td>
<td>0.4587</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>20</td>
<td>Equipment Counts are equal for</td>
<td>L.Bound* E.Price</td>
<td>0.3718</td>
<td>Fail to Reject</td>
</tr>
<tr>
<td>21</td>
<td>Average Link Distances are equal for</td>
<td>Meters</td>
<td>0.0001</td>
<td>Reject</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ($H_0$)</th>
<th>$Prob &gt; t$</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Average Link Distances are equal for Area</td>
<td>0.0000</td>
<td>Reject</td>
<td>Effect is statistically</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>23</td>
<td>Average Link Distances are equal for L.Bound</td>
<td>0.7827</td>
<td>Fail to Reject</td>
<td>Not statistically</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>24</td>
<td>Average Link Distances are equal for E.Price</td>
<td>0.0145</td>
<td>Reject</td>
<td>Effect is statistically</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>25</td>
<td>Average Link Distances are equal for Meters*</td>
<td>0.0004</td>
<td>Reject</td>
<td>Effect is statistically</td>
</tr>
<tr>
<td></td>
<td>Area</td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>26</td>
<td>Average Link Distances are equal for Meters*</td>
<td>0.2220</td>
<td>Fail to Reject</td>
<td>Not statistically</td>
</tr>
<tr>
<td></td>
<td>L.Bound</td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>27</td>
<td>Average Link Distances are equal for Meters*</td>
<td>0.1822</td>
<td>Fail to Reject</td>
<td>Not statistically</td>
</tr>
<tr>
<td></td>
<td>E.Price</td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>28</td>
<td>Average Link Distances are equal for Area*</td>
<td>0.5890</td>
<td>Fail to Reject</td>
<td>Not statistically</td>
</tr>
<tr>
<td></td>
<td>L.Bound</td>
<td></td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>29</td>
<td>Average Link Distances are equal for Area*</td>
<td>0.0772</td>
<td>Fail to Reject</td>
<td>Not statistically</td>
</tr>
<tr>
<td></td>
<td>E.Price</td>
<td></td>
<td></td>
<td>significant</td>
</tr>
</tbody>
</table>

Continued on next page
A discussion of findings is provided in the next section. However a quick look at the above table shows that for the data set used (the scenario and input levels used) Meters, Area, and Equipment-Price, (but not Lower-Bound) are factors that influence Total Network Cost. Also, Meters alone influence Equipment Counts. And lastly, similar to Total Network Cost, Meters, Area, and Equipment-Price influence Average Link Distance.

### 3.11. Findings

Prior sections provided test data and statistical analysis of the experimentation performed using IFNET on a suburban-type set of experimental runs to obtain optimal network solutions. This set of network design instances were representative of suburban areas.

This section summarizes the findings of the experimentation for the three key characteristics of the networks: the Total Network Cost (Total Network Cost) of the networks generated by the experiment, the Equipment Count of these networks, and the Average Link Distance of the links comprising these networks. These three key network characteristics are the outputs of the design process using IFNET. They are influenced by the inputs to, and conditions of, the design instances. The findings here focus on the practitioner implications of the experimentation and study.

---

**Table 3.86 – Continued from previous page**

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothesis ($H_0$)</th>
<th>$\text{Prob} &gt; t$</th>
<th>Conclusion</th>
<th>Which Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Average Link Distances are equal for L.Bound* E.Price</td>
<td>0.7864</td>
<td>Fail to Reject</td>
<td>Not statistically significant</td>
</tr>
</tbody>
</table>
3.11.1. Total Network Cost

Practitioners are typically given a project scope such as area to service, type of service to design and a budget to work within. With these conditions the network designer has a keen interest in the Total Network Cost of the resulting design.

Cost minimization is critical to provide the most service capability, at the quality required (i.e., reliability, security, maintainability), and at the lowest cost possible. The results of this experiment suggest that focusing on the first factor, Equipment-Price, is important in lowering Total Network Cost. The price is a design decision for what equipment to purchase. However, given that pricing of equipment in a competitive environment may not be much different among the choices available, and that it takes an order of magnitude price difference to double Total Network Cost, there likely is not much a designer can do to lower Total Network Cost through focus on Equipment-Price alone.

The second factor important in the design of a network is the (geographic expanse of) area of the project, which is typically a given condition of the project. Total Network Cost is lowered by a reduction in the Area to be designed. The practitioner may or may not be in a position to influence the scope of this aspect of a major project. However, to know that Area is high on the list of factors that influence Total Network Cost, is important to give engineering management a perspective of what is driving Total Network Cost.

Another condition that drives Total Network Cost is the quantity of meters in the project. This also is a condition of most projects, but can be influenced by practitioners. For example reduction of meters can take place if the project is scoped only to address residential meters versus commercial meters. Also, a class of residential meters, such as single family homes versus apartment complexes, can be chosen for first deployment. These decisions have an effect on Total Network Cost through the factor Meters.
Interestingly, the fourth factor, Lower-Bound, did not show a statistically significant effect on Total Network Cost, for the ranges of factors in this experiment. Perhaps by setting the Lower-Bound significantly higher than the present settings, a statistically significant effect might be obtained.

Thus, Total Network Cost is positively correlated to increases in Equipment-Price, Area and Meters. These three factors play a key role in Total Network Cost. Awareness of the above factors’ influence on Total Network Cost allows the practitioner to enter design projects with proper focus on the most relevant conditions of the project and design decisions to be made. Interestingly, designers may not obtain significant lowering of Total Network Cost by focusing on lower bounds for equipment, for the types of conditions given here.

3.11.2. Equipment Count

The practitioner setting out to design a smart-meter project has an interest in minimizing the number of candidate sites brought into service for a network design. If a network can be designed with minimum sites, the design would translate into fewer assets to maintain and in some cases lower recurring real estate costs to fund on an ongoing basis. The experimentation focused on the Equipment Count, or units of equipment (the sum of number of repeaters and collectors), in a solution. Each equipment unit would require a site for placement of the equipment, and thus, each equipment-count drives a site-count, an important aspect of the cost of a network design.

Interestingly, only the factor Meters statistically significantly influence Equipment Count. The influence is positive, in that an increase in Meters correlates to an increase in Equipment Count. The practitioner at times has some influence in modifying the number of meters in a project, but many times this is a given in the smart-meter project that cannot be changed. Thus, site reduction is a function, primarily of other design decisions or business decisions rather than area, equipment price or lower
bound, for projects that closely match the nature of the experiment conducted here. Again, for the ranges of factors studies, the design decision of Lower-Bound did not significantly have an effect on Equipment-count, and thus site count.

3.11.3. Average Link Distance

For practitioners designing networks, Average Link Distance is an important output characteristic of a network design solution, since it has an effect on a key performance metric, namely, link reliability. The longer a communications link, particularly a wireless link, the less signal strength, and ultimately the less reliable the link. The reliability of a link, as a function of link distance, is highly nonlinear. Increases in distance matter little for short links, but matter greatly as the links stretch out to the limits of the technology. This metric is increasingly important when the Average Link Distance comes close to the range limit of the equipment.

This experiment found that the factor Area has an effect on Average Link Distance. The more Area in the project, the higher is Average Link Distance. Also, the fewer Meters in the project the higher the Average Link Distance. Lastly, the higher the factor of Equipment-Price, the higher the resulting Average Link Distance. It was noted that the factor Meters has an effect on Average Link Distance in the opposite direction. The more Meters the lower the Average Link Distance. Interestingly, the interaction factor of Meters*Area, which is representative of meter density, significantly has an effect on Average Link Distance. The lower the density, the higher the Average Link Distance. For practitioners, this suggests, for conditions matching the present experiment, that the larger the Area and the fewer the Meters, the more focus should be placed on Average Link Distance. In other words, as the area becomes vast and meters sparse, Average Link Distance may increase to the limits of the technology. This is precisely the situation in rural and remote areas; these regions are difficult to find a solution for, given high Average Link Distances. Conversely, as density increases, Average Link Distance decreases, as expected. Thus experimentation confirms that the factor Area and density, as proxied by Meters*Area, are the
primary factors that have an effect on Average Link Distance.

3.11.4. Summary Findings

The preceding experimentation was performed to investigate solution characteristics obtained by the IFNET solver for sixteen instances of a suburban deployment scenario selected for study in this research. The iterations represent possible conditions and practitioner-decisions that are part of this practical scenario.

When given a smart-meter project to plan and design, practitioners are faced with the conditions of the project, such as the size of the area and the quantity of meters to deploy. They are also faced with design decisions to be made, including setting of performance parameters, such as a lower bound on traffic through equipment. They must also take a hard look at the price to be paid for the equipment.

The experiment found that the cost of the network can be minimized by practitioner-focus on equipment price, scope of the meters deployed, and area to be deployed. By minimizing these factors, Total Network Cost is reduced. To improve reliability of the network, through a reduction of Average Link Distance, the practitioner should focus on, to the extent possible, increasing the project scope of meters and decreasing the project scale of area, effectively increasing the density of the meter deployment. The reduction of equipment count, which brings indirect cost reductions, such as site fees and maintenance costs, is best done through the judicious reduction of the project scope in terms of meter count. Lastly, for the design decision of lower bound on equipment traffic, there is no statistically significant impact on reducing Total Network Cost, Equipment Count, nor Average Link Distance, at least for the scenario solved here – a counterintuitive result.

The experimentation here illuminates the degree and relative importance of several factors that the designer must be aware of to affect the outcome of design in terms of Total Network Cost, Equipment Count, and Average Link Distance. These responses were shown to have an effect on them by Meters, Area, Equipment-Price and Lower-Bound in some obvious and sometimes interesting and counterintuitive ways.
Chapter 4

CONCLUSIONS AND FUTURE RESEARCH

This research applied an advanced optimization technique, interval-flow, and its instantiation in software, IFNET, to a practical industry problem. A mathematical model, SG1 was formulated, and applied to several instances of the smart-grid communications network design problem to arrive at a set of solutions. SG1 model equations (2.22)-(2.32) appear on page 53. Experimentation addressed several conjectures regarding how certain key network design figures of merit responded to various design decisions and conditions. Conclusions were drawn for this specific set of experiments.

4.1. Summary of Findings and Conclusion

IFNET shows promise as a way to quickly solve large problems that can benefit from speed and special treatment of lower bounds on decision variables, in this case on equipment capacity. The study found that IFNET can be applied successfully to smart-grid problems, and the solver was successfully applied to a subset of the smart-grid problem – the smart-meter problem for suburban areas. The findings include how the practitioner-designer can influence key characteristics of a proposed network solution, through the judicious influence on factors leading up to the design work, and on the design decisions the designer can make during the project.

The IFNET solver when applied to a suburban scenario, through an experiment that tested a range of network solutions, shows promise in allowing practitioner-designers the ability to influence key business and engineering outcomes. It does so by allowing awareness of the impact of project situational conditions such as size and scope of the project, and by the ability to make project design decisions that may
influence total network cost, site requirements such as equipment count, and network quality metrics such as link reliability arising from average link distance.

4.2. Contributions of the Research

This research is important for solving the problem of efficiently and effectively designing smart-grid communications networks. Smart-grid communications networks are now being deployed in rural areas. Urban areas are now considering a second generation of technology. Many new proprietary technologies are being proposed, and some standardization is being adopted. Much of this planning is being done without the benefit of network topological optimization, nor the advanced new optimization methods and software, such as IFNET, now emerging for such use. This leads to sub-optimal designs for these large programs. This research demonstrates the value of a breakthrough optimization solution technology that scales well to large projects and is eminently viable in reducing implementation costs of the smart-grid networks.

This research provides a contribution to industry by applying an advanced optimization tool from academic research to the important industry problem of smart-grid. Smart-grid programs cost utilities, and ultimately society, hundreds of millions of dollars. Application of an advanced optimization technology and solver from academic research holds promise for achieving significant savings to industry in the design and deployment of these networks. This work has contributed to advancing this research in academia and proposing a solution methodology for network design problems for industry. The following list includes some of the contributions of this study.

- Use of an advanced research heuristic for quickly solving mixed integer linear programs, is applied to a practical utility problem, the smart-grid communications network design problem.[5]

- The IFNET solver was successfully applied to a key subset of the smart-grid
problem, the smart-meter problem.

- The experimentation tested important input (factors) and output (responses) of the design process for an important sub-class of the smart-meter problem, the suburban zone smart-meter deployment problem.

- The study quantified the influence of key factors to key responses, thereby uncovering the influence of each factor individually, joint factors in combination, and all factors studies in relation to one another.

- The study allows practitioners to give emphasis to factors which may drive design figures of merit that are important during strategy and planning.

- The study technique allows both technical and financial aspects of the problem to be simultaneously solved.

- The approach taken here could potentially save utilities millions of dollars of cost to rate-payers (citizen-consumers in society).

4.3. Future Research

Smart-grid networks, one of the first examples of the “Internet of Things” (IoT) will evolve to encompass millions of devices that require sophisticated communications and control. Therefore the model and algorithmic approach started in this research should be extended in scope to handle these millions of devices, perhaps via use of cloud-based parallel processing techniques.

These industrial IoT networks will require ever-improving prescriptive and predictive models to assist the analytics processes in the network perform important tasks such as self-healing, predicting failure events, anticipating and reacting to consumer demand, and scheduling preventative maintenance. This research provided one deterministic prescriptive model for optimizing the network topology, and one statistical
predictive model for predicting responses to input factors. Variants on these models can be provided for a multiplicity of needs in the coming decades, and deployed across the system in embedded edge device software, area gateways, and in the core “big data” IoT platforms.

Comparison of this research’s results to actual SG network designs was problematic due to network security and customer confidentiality issues. Future work is suggested, from within utilities, toward comparing the plans and results of current SGCN deployments with results from this tool and experimentation methodology.

The present state of the network generator and solver tool allows for entry of thirteen parameters per run (experimental observation). Future research may take advantage of automation of the tool’s data input capability, such that a larger number of feasible solutions can be generated, thereby allowing for the investigation of a wider range of scenarios beyond neighborhood density, for example, towards a study of urban and rural deployments.

This study solved problems based on idealized network topologies. Future studies could take into account complexities of real networks including the need for line-of-sight links for wireless technologies, the need to accommodate obstacles and rights-of-way for wired technologies (fiber, copper, and coax), and a broader setting of network access configurations (beyond point-to-point, to fully meshed, and point-to-multipoint capabilities for example). Because designers of smart-grid networks use such a wide variety of communications and networking technologies to realize smart-grid network designs — such as wireless, cellular, microwave and fiber infrastructure — the method used here could evolve to include handling particular constraints posed by each of these technologies. Wireless channels in particular require incorporation of accurate terrain and clutter databases into a 3D surface model of the Earth to accurately represent the characteristics of the radio channel. This is an active area of research at the crossroads of the disciplines of geographic information systems and radio propagation prediction within electrical communications engineering.
Finally, with the ability to analyze immense amounts of input data comes the risk of generating massive amounts of output (solution) data that needs to be reduced to useful information for the researcher and practitioner. Visualization techniques help convey this massive amount of solution data into easy to grasp intermediate and final answers. Research is indicated into effective techniques for generating network graphs or maps or network links over satellite images of terrain and clutter, to thereby visualize and interpret the topological network solution (connectivity) and solution value (overall network cost). The author is actively pursuing this as post-doctoral future research, with potential commercial applications.
Bibliography


[23] Department of Energy (2010). *Communications Requirements of Smart Grid Technologies*. DoE.


189


190


