Human Inspection Variability in Infrastructure Asset Management: A Focus on HVAC Systems

Clayton Michael Pratt

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Civil Engineering

Tripp Shealy, Chair
Farrokh Jazizadeh
Michael Garvin

13 December 2023
Blacksburg, VA

Keywords: Infrastructure Asset Management; Inspector Uncertainty; Facilities Management; Inspection Variability; Sustainment Management System; HVAC; Predictive Modeling

Human Inspection Variability in Infrastructure Asset Management: A Focus on HVAC Systems

Clayton Michael Pratt

Abstract

Human inspection is a pivotal component of infrastructure asset management within a systems thinking approach to civil engineering. Skilled inspectors are tasked with the evaluation of various civil infrastructure components, conducting assessments of their conditions, identifying maintenance needs, and determining necessary repairs. Despite the growing interest in advanced technologies and automated inspections, the use of human-in-the-loop procedures is still widely practiced. Humans are susceptible to cognitive bias, variability, or uncertainty when inspecting infrastructure, and finding solutions to reduce these factors is paramount.

This study presents a comprehensive exploration of inspection variability within infrastructure asset management, drawing insights from datasets of the BUILDER Sustainment Management System (SMS) program. The research delves into infrastructure inventory, inspector data, and inspection data components of an asset management database, shedding light on variability in human inspection. Variations in inspection ratings revealed significant concerns, particularly in Mechanical, Electrical, and Plumbing (MEP) systems, with notable disparities between inspection ratings and condition ratings. Inspector variability analysis, through Coefficient of Variation calculations, indicated substantial disparities within and among inspectors. Further analysis, including Tukey's

HSD test, pinpointed significant variability in heating, ventilation, and air conditioning (HVAC) and Fire Protection system inspections.

Moreover, this study addresses the specific challenge of reducing inspection uncertainty in HVAC systems. HVAC systems play a critical role in facility energy consumption, and their maintenance is vital to energy efficiency and occupant comfort. However, HVAC-specific inspections primarily require human involvement, making them time-consuming and prone to error. Addressing the challenges surrounding human inspection of HVAC systems, this research presents a multifaceted approach to reduce variability. Drawing from a review of existing literature on HVAC inspection uncertainty, this study extends its focus to the development of predictive models. These models considered parameters including inspection ratings, age-based obsolescence, section condition indices, component characteristics, and unique inspectors. Utilizing Linear Regression, Random Forest, and Gradient Boosting Regression, this model accurately predicted Variability Ratings, signifying the potential for implementation as a decision support tool. Importantly, the findings highlight the need to not only understand the factors affecting HVAC inspection variability but to actively implement technological solutions that can reduce human error and variability in inspections.

Human Inspection Variability in Infrastructure Asset Management: A Focus on HVAC Systems

Clayton Michael Pratt

General Audience Abstract

Infrastructure inspection is crucial for maintaining buildings and facilities, but it often comes with human errors and uncertainties. This study looks at the inspection process, focusing on case studies and data from the BUILDER Sustainment Management System (SMS) program. It reveals that inspectors sometimes evaluate the condition of parts of a building differently, leading to inconsistencies and poor overall management.

One significant area of concern is heating, ventilation, and air conditioning (HVAC) systems. These systems play a critical role in facility energy use and can be challenging to inspect accurately. Previous research has shown that work experience, training, education, and other factors tend to contribute to variability in how inspectors assess HVAC systems. This research not only highlights these issues but also develops predictive models to reduce the variability of HVAC inspections. By doing so infrastructure can be managed correctly and ultimately lead to improved building lifecycles.

Acknowledgements

I am deeply grateful for the invaluable support and guidance I received throughout this journey. I would like to express my heartfelt appreciation to Dr. Tripp Shealy and Dr. Farrokh Jazizadeh, my co-chairs, for their unwavering patience, kindness, and dedication. Both Dr. Jazizadeh and Dr. Shealy not only helped me navigate the complexities of academia but also nurtured my sense of curiosity and wonder, igniting a passion for research and scholarship. Their mentorship played a pivotal role in the completion of this Master's thesis, and I look forward to continuing this scholarly voyage as I move to a teaching position at the Air Force Institute of Technology. I am excited about the growth and possibilities that lie ahead.

I'd also like to extend my gratitude to my committee member, Dr. Michael Garvin. Dr. Garvin's wisdom, both in and out of the classroom, provided invaluable insights into conducting my research, maximizing learning in the VCEMP program, and growing as a military officer.

Lastly, I cannot thank my family enough. My wife, Audra Beth, has been a constant source of support throughout this program, all while pursuing a career and graduate degree of her own. My father and mother, Mike and Lipi, instilled in me the importance of education and continuous learning. My brother, Heath, has taught me to find the positive in every situation. My grandfather, Bharat, inspired me to be an engineer many years ago.

Table of Contents

List of Figures

List of Tables

Thesis Introduction

Human inspection is a critical element in infrastructure asset management, with skilled inspectors playing a pivotal role in assessing and evaluating civil infrastructure components, identifying maintenance needs, and determining necessary repairs (Agnisarman et al. 2019). While the integration of advanced technologies and automated inspections through computer vision has made significant strides, complete replacement of human inspectors by technology remains elusive, partly due to the intricate nature of infrastructure assets and the difficulty in replicating human expertise (Jing et al. 2022; Sharma et al. 2023; Zhe et al. 2019). Currently, a combined approach has emerged, one that seeks to utilize both humans and technology to enhance infrastructure asset management systems, recognizing the role of human expertise in the process. Identifying the areas of greatest variability in human inspections is paramount to pinpointing the most urgent needs for further research (See et al. 2017).

Human Inspection Problems

Human inspection is often characterized with challenges such as cognitive bias, subjectivity, and variability (Gordan et al. 2022). Within infrastructure asset management human inspection is prevalent, making these challenges important to address. These inherent challenges are further reflected in differences of judgment, evaluation, and decision-making among inspectors, resulting in varying assessment outcomes. Additionally, factors such as work experience, training, inspection duration, and environmental conditions have been identified as key contributors to the observed variability in human inspection findings (Campbell et al. 2021). It is evident that there is a pressing need to address these human inspection problems in order to enhance the accuracy and reliability of infrastructure assessments.

Need for Identifying Inspection Variability

Recognition of variability within human inspection processes is a fundamental prerequisite for improving infrastructure asset management systems. In this context, this study utilizes datasets from the Department of Defense's infrastructure asset management program, to quantify, pinpoint, and reduce significant inspector variability. The research employs the BUILDER Sustainment Management System (SMS), a web-based tool that facilitates real-time data management and assigns condition indices to infrastructure assets based on their expected lifecycles ("BUILDER Sustainment Management System" 2012). However, despite its utility, the system faces challenges due to the high level of expertise required for its effective use, organizational constraints, and uncertainties regarding cost savings (Herrera et al. 2017). The need to improve the current system and reduce inspection variability is underscored by the pressing issue of misaligned asset management investments, particularly in the context of the Air Force's substantial infrastructure deferred maintenance backlog (Synovec et al. 2023).

Need for Reducing Inspection Variability

The final step of this research lies in quantifying and mitigating human inspection variability, with a specific focus on HVAC systems, which serve as a critical testing bed. This study is particularly significant within the context of infrastructure asset management, especially within the Air Force. The selected emphasis on HVAC systems is driven by their inherent importance to facilities, contributing significantly to energy consumption and operational costs (Cai et al. 2023; Tharanga et al. 2022). By building upon the foundational concepts reviewed in earlier research, this study delves deeper into the challenge of reducing inspection variability in the context of these complex systems. The objective is to harness the insights and lessons learned from previous work to develop practical strategies and a robust model capable of effectively addressing the issue of variability within HVAC inspections. This endeavor ultimately contributes to the improvement of infrastructure asset management systems, ensuring more efficient and precise allocation of resources for the maintenance and preservation of critical infrastructure assets (Zhe et al. 2019).

Summary of Results

This research is focused on identifying, quantifying, and reducing variability in human inspection of infrastructure. Inspection of HVAC systems was ultimately chosen as an area of emphasis, but the applications are broadly applicable to all building and infrastructure systems. Additionally, this research is positioned as a starting point and stepping-stone for future research in the areas of infrastructure asset management and facilities management. The major research questions and results from this study are shown in **Figure 1**, below.

Infrastructure Inspection Questions

What variability exists within human inspections processes of the BUILDER SMS program?

How much variability exists within human inspection processes of the BUILDER SMS program?

Infrastructure Inspection Results

Variability exists between inspectors and by inspector, particularly in MEP systems.

Coefficient of Variation values from 0.12 to 6.42 with 93% of inspectors >1. Shows variability between and among inspectors.

HVAC Inspection Questions

What strategies exist to reduce human inspection uncertainty in HVAC systems?

Can a model be developed to reduce inspection uncertainty in HVAC systems?

HVAC Inspection Results

Automated technology, predictive modeling, training, education, demonstrating tasks.

Regression Models can accurately predict a Variability Rating to identify potential high variability inspections.

Figure 1. Research questions and results.

Journal Paper 1: Inspection Variability for Infrastructure Asset Management: A Case Study of the BUILDER SMS Program

Inspection Variability for Infrastructure Asset Management: A Case Study of the BUILDER SMS Program

Abstract

Human inspection is a pivotal component of infrastructure asset management. Skilled inspectors are tasked with the evaluation of various civil infrastructure components, conducting assessments of their conditions, identifying maintenance needs, and determining necessary repairs. Despite the growing interest in advanced technologies and automated inspections, the use of human-in-the-loop procedures is still widely practiced. Given that humans are susceptible to cognitive bias, variability, and uncertainty when inspecting infrastructure, understanding the intricacies of human involvement is critical. This aligns with the focus of this study, which investigates infrastructure inventory, inspector data, and inspection data components of an asset management database. By delving into these areas, this study explores human inspection processes to identify and quantify inspector variability. A secondary data analysis using qualitative and quantitative statistical processes was tested on Air Force infrastructure datasets from the BUILDER Sustainment Management System (SMS) to reveal insights into infrastructure inspection practices. Variations in inspection ratings highlighted notable disparities between inspection ratings and condition ratings, particularly in Mechanical, Electrical, and Plumbing (MEP) systems. Inspector variability analysis, through Coefficient of Variation calculations, indicated significant disparities within and among inspectors. Finally, Tukey's HSD test identified significant variability in HVAC and Fire Protection system inspections. Ultimately the presence of human inspection variability was

confirmed and paves the way for further research to reduce these discrepancies in specific types of infrastructure assets.

Introduction

Human inspection of the built environment holds significant importance within the domain of infrastructure asset management. This human inspection involves skilled inspectors assessing components of civil infrastructure systems, evaluating the condition, identifying maintenance requirements, and assessing repair needs (Agnisarman et al. 2019). Following inspection, inspectors input evaluations into facilities management or asset management systems, forming an essential part of a human-in-the-loop approach. The integration of advanced technologies and automated inspections through computer vision has garnered attention and investment, but a stage where human involvement can be entirely replaced has not been reached (See et al. 2017; Zhe et al. 2019). One major challenge to this is that not all assets are created equal, and the reliability of technology to match the expertise of human inspectors is difficult (Jing et al. 2022; Sharma et al. 2023; Zhe et al. 2019). Instead, a pragmatic approach has emerged - one that advocates for the synergy between humans and technology through infrastructure asset management systems that complement human inspectors rather than replace them. Understanding where this human inspection exhibits the greatest variability is paramount to pinpointing the most urgent needs for further research. Previous studies have suggested that improvements in training, inspection procedures, and equipment can help reduce this variability (See et al. 2017). More importantly, this insight paves the way for the initial steps in designing systems that account for the inherent human behaviors and biases entwined within the inspection process.

This study investigates the human inspection processes within a facilities management database, to identify and quantify inspector variability. This is key in providing insight into potential human inspection improvements for civil infrastructure systems. By highlighting areas where human inspection exhibits variability, future research can begin to develop more effective, human-inclusive asset management systems, that reduce asset variability. The Background section provides an overview of the BUILDER SMS program and the research questions to be answered. The Methods detail the process used to answer the research questions. Results and Discussion offer findings based on the methodology used. The conclusion provides a summary, future research opportunities, and implications of these findings for industry.

Background

This research focuses on the inspection element of previously established asset management framework and its use for assessing infrastructure condition to inform decision-making (Thomasson and Sinha 2015). Visual inspection is a prevalent technique for evaluating infrastructure condition, however there is significant research into new technologies to remove the human element from this process (Kim et al. 2022; Moradi et al. 2019). Human conducted visual inspection has been shown to include variability due to differences in human judgment and evaluation of data, often referred to as cognitive bias (Gordan et al. 2022). For example, an experiment on a bridge system showed distinct differences between a group of 96 inspectors in identifying structural defects (Liu et al. 2023). Human inspection is also labor intensive and prone to decision biases, meaning different inspectors looking at the same infrastructure may draw different conclusions (Gordan et al. 2022) Factors such as work experience, training, inspection duration, and environmental conditions contribute to variability in human inspection findings (Campbell et al. 2021). Inspectors with varied experiences and comprehensive training tend to conduct inspections consistently and make informed decisions. The duration and environmental conditions of an inspection can influence assessment thoroughness, potentially causing variations in infrastructure condition identification and evaluation (Campbell et al. 2021).

This research used human inspection datasets within the Department of Defense's (DoD) infrastructure asset management program, with specific regard to the Air Force. With 180 installations worldwide and over $263 billion in infrastructure, the Air Force is a major federal organization that has identified the need for proactive, data-drive asset management (Wilson and Goldfein 2019). The DoD currently uses the BUILDER Sustainment Management System (SMS) for facility asset management that was created by the U.S. Army Engineer Research and Development (ERDC) in 2009 ("BUILDER Sustainment Management System" 2012). This web-based manages real-time data including asset conditions, costs, inspections, renovations, material information, and equipment information. BUIDLER SMS assigns a condition index to all infrastructure assets based on an expected lifecycle and provides feedback on repair, rehabilitation, or replacement as a Decision Support System (DSS) ("BUILDER Sustainment Management System" 2012). All assets contained within BUILDER SMS are categorized and organized by infrastructure type using the ASTM UNIFORMAT II building element classification scheme. The BUILDER SMS platform is predicated on Knowledge Based Inspections (KBI), which determines for a user what infrastructure assets need an inspection and when. The KBI differ from a typical regularly scheduled maintenance and inspection plan (Uzarski et al. 2007). The KBI are determined by using a Performance-Failure (P-F) curve that estimates the deterioration of an asset through its lifecycle as seen in **Figure 2**, below (Uzarski et al. 2007, 2018). Based on this curve there is an identified "Sweet Spot" range at which investing in maintenance or repair of a system will maximize the lifecycle of an asset (Uzarski et al. 2007). For the Air Force, this range is from a condition index of 60 to 80 on a scale from 0 to 100 (Lamm et al. 2022). According to this principle, once an asset has passed its Remaining Maintenance Life (RML) it is no longer cost-effective to invest in the asset and it should then be run-to-failure. By performing KBI, the P-F

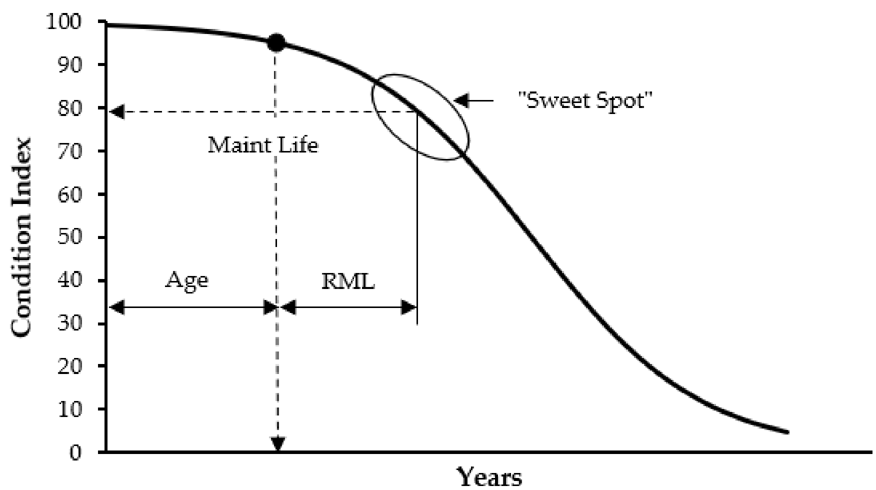

Figure 2. P-F curve used for expected lifecycle deterioration of infrastructure by BUILDER SMS. Adapted from: (R. et al. 2007)

curve can be refined and tailored to a specific infrastructure system as opposed to following this generic curve ("BUILDER Sustainment Management System" 2012).

There is research to show limitations with the BUILDER SMS program. While KBI provides cost-savings by only focusing on mission-critical and condition-critical assets, it requires a high level of engineering expertise and experience to be done correctly (Bartels 2014). Some literature has suggested remote sensors, optical interferometry, spectral analysis, or thermal infrared analysis as better alternatives to the current KBI (Bartels 2014). Other research has shown that BUILDER

SMS has presented several challenges with integration into the DoD such as organizational buy-in, uncertain cost savings, and information assurance constraints (Herrera et al. 2017).

Another concern with BUILDER SMS system is the standardized P-F curve that is used to describe the lifecycle aging of a given infrastructure asset. Historical condition index data of 8,549 roofs across the DoD were modeled using stepwise functions and shown to be more accurate than the current BUILDER SMS model 92% of the time (Lamm et al. 2022). In addition, this research included historical condition index data rather than a manufacturer recommended service life, which is what BUILDER SMS currently employs (Lamm et al. 2022).

The DoD recognized issues with variability caused by human inspection and has issued a standardized guideline for conducting inspections in an effort to reduce differences (Uzarski et al. 2018). Given the fact that the Air Force's infrastructure deferred maintenance backlog has continued to grow to over $33 billion in Fiscal Year 2022, there are still accuracy concerns with human inspections driving misaligned asset management investments (Synovec et al. 2023). With a lack of investment by the Air Force to remove human inspections, there is a need to improve the current system and reduce variability, ensuring money is invested in the right infrastructure at the right time.

The objective of this research is to quantify variability within human inspection components of the BUILDER SMS infrastructure asset management program. With this objective the primary research questions were:

1. What variability exists within human inspections processes of the BUILDER SMS program?

2. How much variability exists within human inspection processes of the BUILDER SMS program?

To answer research question one an analysis of inspector data was conducted. By qualitatively showing differences between inspector's evaluations of infrastructure variability could be confirmed and located. To answer research question two an analysis of inspection data was conducted. By looking at differences in individual inspections, statistical analysis could be used to quantify variability and pinpoint areas for future research.

Materials and Methods

To answer the proposed research questions primary data from BUILDER SMS was collected from Air Force Bases across the United States, shown in **Figure 3**, below. Overall data analysis was completed in three distinct areas, infrastructure inventory, inspector data, and inspection data. Each of these are further explained in the following subsections.

Figure 3. CONUS Air Force Base names and locations.

Infrastructure Inventory Analysis

Infrastructure inventory data collection involved generating several pre-existing reports from BUILDER SMS over a collection period from February to May 2023. A Building System report was generated on 57 active-duty Air Force bases in the Contiguous United States (CONUS) to provide data on the entire infrastructure asset inventory and currently assessed building conditions. This dataset included an inventory of 110,335 unique infrastructure assets. Information on the assets provided in this report is included in Appendix A. All reports were combined into a working data frame that was used for data analysis in a Juptyer Notebook with Python coding.

Initial analysis of the infrastructure inventory dataset involved examining correlations between established metrics such as Building Condition Index (BCI), Year Built, and Mission Dependency Index (MDI). This correlation analysis was similar to one conducted on the relationship between BCI and Building Functionality Index by McDonald (2023). This was necessary to see if inspection data trends could be attributed to underlying relationships with the infrastructure itself. Comparisons of Year Built and MDI to BCI were initially conducted as heatmaps to visualize concentrations of data for trends and relationships. This was then followed by calculating Spearman's Rho values for correlation. These distributions included all available data and only excluded data where a value for BCI, Year Built, or MDI was not included. This resulted in 45,420 data points used for visualization and analysis.

Additional information on the infrastructure inventory was also collected by calculating summary statistics based on UNIFORMAT II building element classification. The Building System report was initially grouped by each building element classification, ranging from A10 to H50. From there data was also grouped into corresponding Air Force Major Commands (MAJCOMs).

categories. These MAJCOMs represent different mission sets and geographic areas across the United States and include labels such as Air Force Mobility Command and Air Combat Command. At this point the mean condition index for all data assigned to an element classification and MAJCOM was taken. To remove outlier sensitivity, any element classification that had condition indices for less than 5% of the population was not included.

Inspector Data Analysis

Data collection on inspectors involved generating several pre-existing reports from BUILDER SMS over a collection period from February 2023 to May 2023. An Inspector Matrix report was generated on 57 active-duty CONUS Air Force Bases. This dataset included an inventory of 20,705 unique inspectors with associated green, amber, and red inspection ratings on different UNIFORMAT II element classifications. Information on the data provided in the Inspector Matrix report can be found in Appendix A. All reports were combined into a working data frame that was used for data analysis in a Juptyer Notebook with Python coding.

For this analysis, the percentage of green, amber, and red scores was averaged at the MAJCOM level and then compared to mean values for condition indices found during infrastructure inventory analysis. This information was plotted to visualize trends, outliers, and significant variability between inspector evaluations and the infrastructure data. The range of condition index values used was 80-100 for green, 60-80 for amber, and less than 60 for red. Additionally, the dataset was classified based on UNIFORMAT II element classification and plotted alongside mean condition indices for similar visualization techniques. Finally, a hypothesis test was conducted quantify how many inspectors' ratings classifications differed from the population mean. This could be used to identify whether variability among inspectors' ratings was present. mean. The following null and alternate were used with a significance of α=0.05. To allow for comparison, a percentage of the number of inspectors that showed significant difference was calculated by MAJCOM.

H_o: There is no significant difference between Inspector X's mean percentage of X (green, amber, red) scores and the population mean of X (green, amber, red) scores.

H_a: There is a significant difference between Inspector X's mean percentage of X (green, amber, red) scores and the population mean of X (green, amber, red) scores.

Inspection Data Analysis

Inspection data collection involved generating several pre-existing reports from BUILDER SMS over a collection period from February 2023 to May 2023. An Inspection Matrix report was generated on 16 active-duty CONUS Air Force Bases to provide data on all inspections that had been completed and reported in BUILDER SMS. Due to the size of this dataset, a random sample of 16 installations was used. This sample ensured that every MAJCOM had at least 2 installations in the sample size to ensure differences between locations and mission sets were captured. This dataset included an inventory of 621,381 unique inspections with associated information on the reported condition rated by an inspector and an expected rating generated by BUILDER SMS. In addition, an Error value was calculated for each inspection as the difference between Inspection and Expected Rating as shown in Equation 1. Information on the data provided in the Inspection Matrix report can be found in Appendix A. All reports were combined into a working data frame that was used for data analysis in a Juptyer Notebook with Python coding.

$$\text{Error}=\left|\frac{\text{Expected Rating-Inspection Rating}}{\text{Expected Rating}}\right| \qquad (1)$$

Initial analysis of the inspection data involved plotting percent deviation of inspection ratings from the expected. For ease of visualization, a stacked bar plot was used, and data was classified by MAJCOM and UNFORMAT II element classification. Additionally, an analysis on mean Error was plotted visually using a bar plot with data being aggregated by UNIFORMAT II element classification and summary statistics for mean error and standard deviation were calculated. All values with an undefined error (Expected Rating = 0) were removed. An analysis was also done by calculating the frequency of inspections (in inspections per year) and plotting versus condition index. To determine the frequency of inspections a count was conducted for each unique asset on the number of inspections included in the dataset. This was divided by the Section Age of the infrastructure asset to provide a frequency value in inspections per year. This data was then plotted against the condition index of that infrastructure section to observe any trends or relationships. The expected relationship was that more frequent inspections would indicate an infrastructure asset that was in poorer condition (deteriorating condition)(West et al. 2022). If this trend was not observed, then a conclusion could be drawn that human inspection variability was a contributing factor.

Finally, a quantitative analysis on inspection variability was conducted. This involved calculating Coefficient of Variation (CV), performing a Kruskal Wallis Test, and conducting a Tukey's Honestly Significant Difference (HSD) Test. A sample of inspectors from the inspection data was taken to reduce computations while still allowing for application to the population at large. The appropriate sample size was calculated using Equation 2, below. The ratio of standard deviation to mean for inspectors, or CV, has previously been used for quantifying differences between inspectors on bridge condition assessments and was adapted for this research (Sein et al. 2019; Washer et al. 2020). The CV was calculated on error values for each inspector and plotted using Equation 3 below. This allowed for observations of variability for specific inspectors and for comparison of variability between inspectors. The Kruskal Wallis Test was conducted as an alternative to an Analysis of Variance (ANOVA) that allowed for non-normality of a distribution (McKight and Najab 2010). This was tested on the inspection data grouped by UNIFORMAT II classification to see if there was a significant difference in mean error between infrastructure types. Based on the Kruskal Wallis Test results, a Tukey's HSD test was conducted to pinpoint specifically which UNIFORMAT II element types differed from each other in terms of mean error, adapted from previous research on error biases in infrastructure condition assessments (Humplick 1992). The results of this test were captured in a spaghetti plot to allow for an easy visualization of inspection variability between infrastructure types.

$$n = \frac{Z^2\sigma^2}{E^2} \tag{2}$$

$$CV = \frac{\sigma}{\mu} \tag{3}$$

Results

Infrastructure Inventory Analysis

The analysis of the infrastructure inventory dataset, depicted in **Figure 4**, reveals non-normal distributions for Year Built, BCI, and MDI values. To explore correlational relationships, Spearman's Rho was computed for Year Built and BCI, as well as MDI and BCI. For the sample size of $n=45{,}420$, the correlation coefficient between Year Built and BCI was $\rho=0.371$, demonstrating a statistically significant weak, positive correlation ($p < 0.0001$). Additionally, the correlation coefficient between MDI and BCI was ρ=-0.065. This is significant with a p-value of less than 0.0001 but a negligible negative correlation (Schober et al. 2018).

Further visual analysis using heatmaps are presented in **Figure 5**. In this instance BCI was a

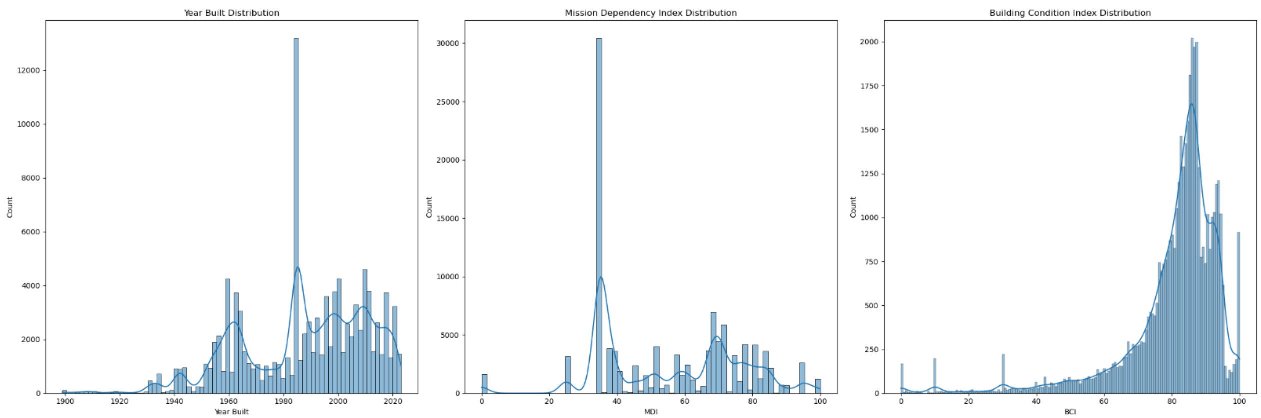

Figure 4. Distribution plots of Year Built, MDI, and BCI variables.

dependent variable plotted against Year Built and MDI. Visualizing the data indicates that Year Built shows a stronger positive correlation than MDI when compared to BCI. Further, MDI has several bands of outlying regions, particularly around an MDI of 25 and an MDI near 0. This may suggest a good lifecycle management plan may be in place even for facilities of low importance.

Inspector Data Analysis

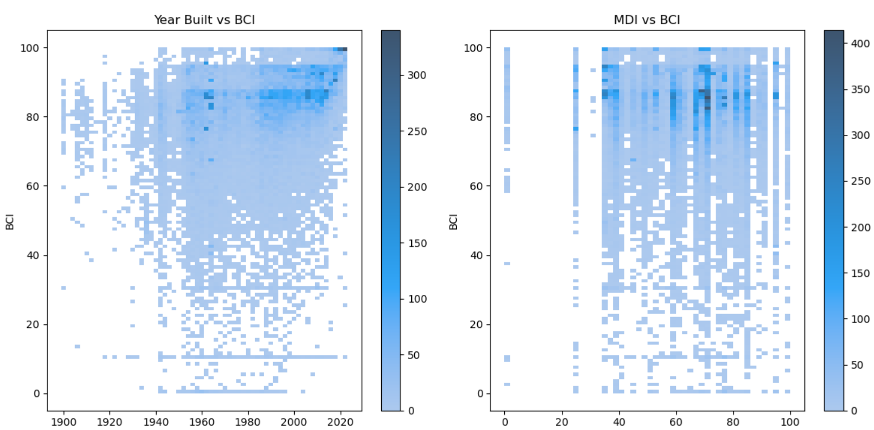

Figure 5. 2x2 square histogram heatmap comparing distribution of: (a) Year Built and BCI; (b) MDI and BCI. Colors indicate counts of data that fall into a specific bin.

The grouped bar plot in **Figure 6** highlights how much variability exists by showing that inspectors tended to evaluate less infrastructure as good condition and more as fair or poor condition when compared to BUILDER SMS condition assessments. For national security purposes, MAJCOM names were left anonymous.

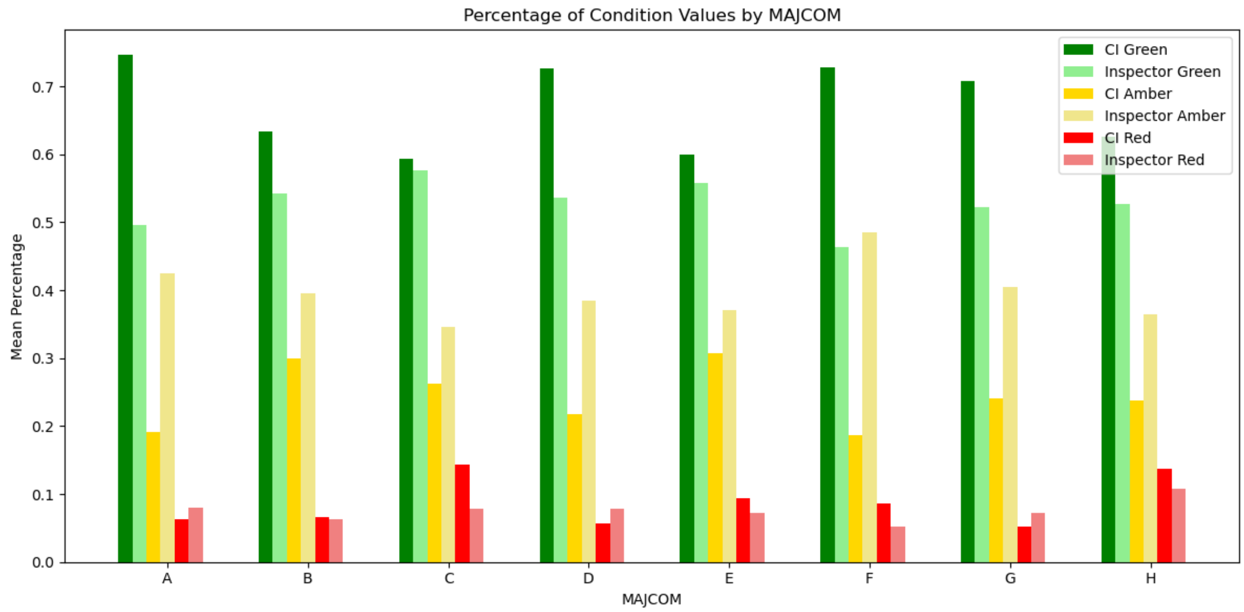

Figure 6. Percentage of Assessed Condition by Inspectors and BUILDER Database.

Further quantifying the amount of variability present is shown in the stacked bar chart in **Figure 7**. This plot clearly illustrates gaps between BUILDER SMS infrastructure conditions and inspector scores with notable discrepancies in HVAC systems. To remove outlier sensitivity, any element classification that had inputs for less than 5% of the population was not included. Results of the hypothesis test on difference between inspector scoring and population scoring are shown in **Table 1**. Notably, all MAJCOMs had varying amounts of significantly different inspectors, with all except MAJCOM D showing high propensity to reject the null hypothesis. This indicates both inter and intra-inspector variability.

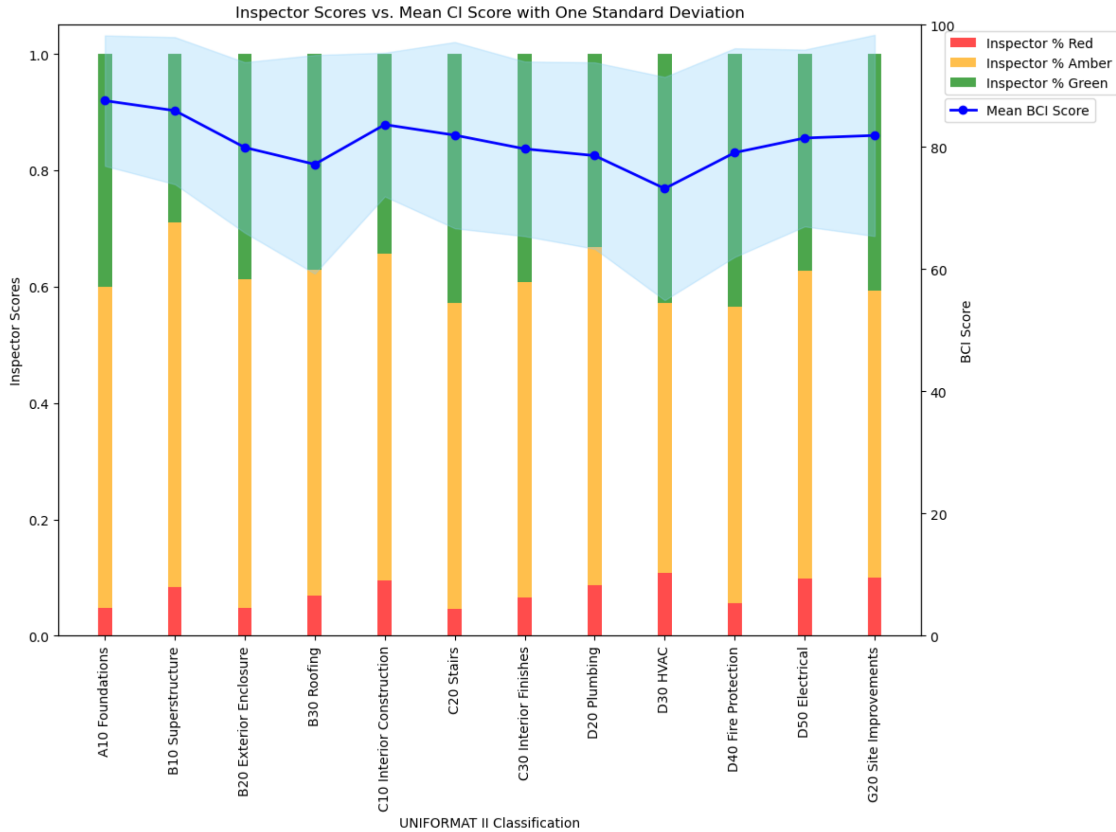

Figure 7. Mean Inspector Scores and Mean CI Scores by UNIFORMAT II Classification.

Table 1. Results of Hypothesis on Percentage of Significant Inspectors from Population Mean.

MAJCOM	% of Significant Inspectors by Green Scores [-]	% of Significant Inspectors by Amber Scores [-]	% of Significant Inspectors by Red Scores [-]
A	27.01	74.34	17.19
B	21.79	65.19	77.67
C	66.80	49.03	49.12
D	0.38	7.43	34.31
E	12.91	68.39	22.82
F	4.19	9.19	8.03
G	88.41	23.45	9.97
H	70.97	26.19	29.59
Average	36.56	40.40	31.09

Inspection Data Analysis

The investigation of research question two using individual inspection data shows high mean error for inter and intra-inspector comparisons and is specifically high in MEP infrastructure systems. Stacked bar plots categorized by Air Force Base and UNIFORMAT II element classification are shown in **Figure 8**. These plots show the percent difference of inspection ratings to BUILDER SMS expected ratings. By Air Force Base distributions are similar with the majority of inspections showing greater than 10% difference from expected ratings. By infrastructure system, plumbing and HVAC exhibit the highest percentage difference in inspections. Installation names were removed and made anonymous for national security purposes. To remove outlier sensitivity, any element classification that had inputs for less than 5% of the population was not included.

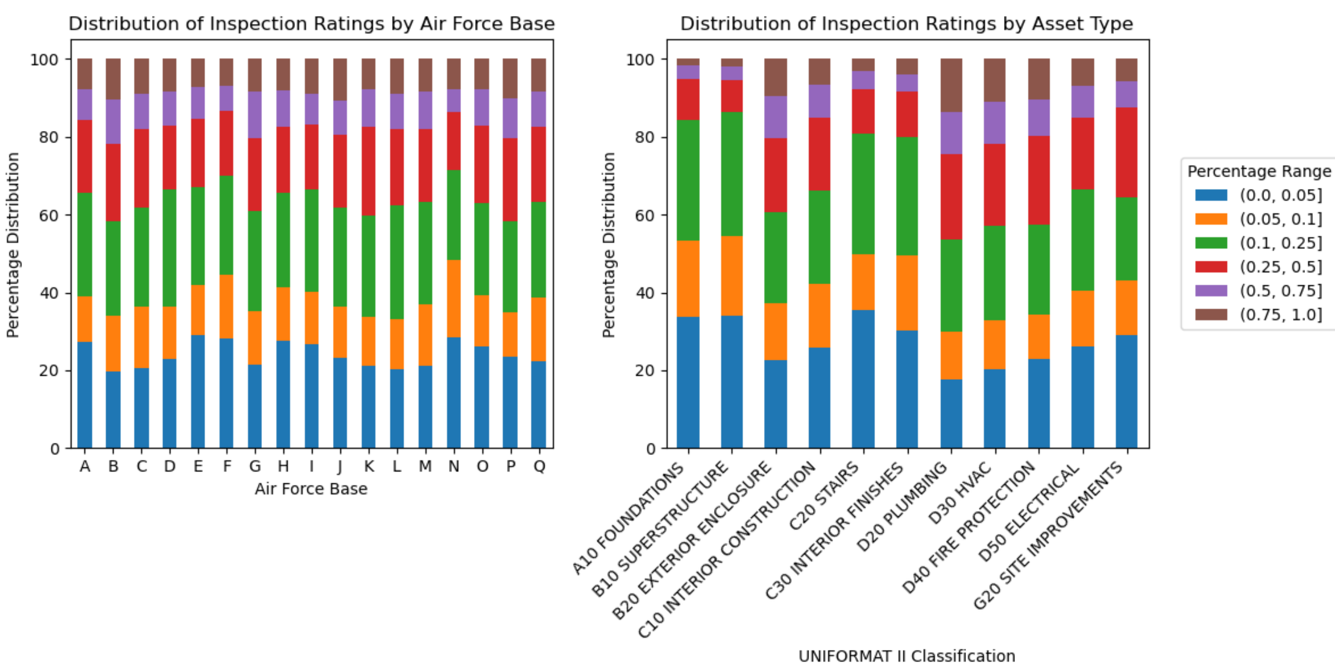

Figure 8. Stacked bar plot of distribution of percentage difference between inspection and BUILDER SMS expected ratings classified by: (a) Air Force Installation; (b) UNIFORMAT II element classification.

Visual and quantitative analysis of mean error by infrastructure system is shown in **Figure 9** and **Table 2**, respectively. Exterior enclosure showed the highest mean error of 6.82 with HVAC, Fire Protection, and Site Improvement infrastructure following closely behind. Generally, higher mean errors also saw exponentially larger standard deviations. All values where error was undefined (Expected Rating=0) were removed, which was 47,048 inspections (7.57%) of the dataset.

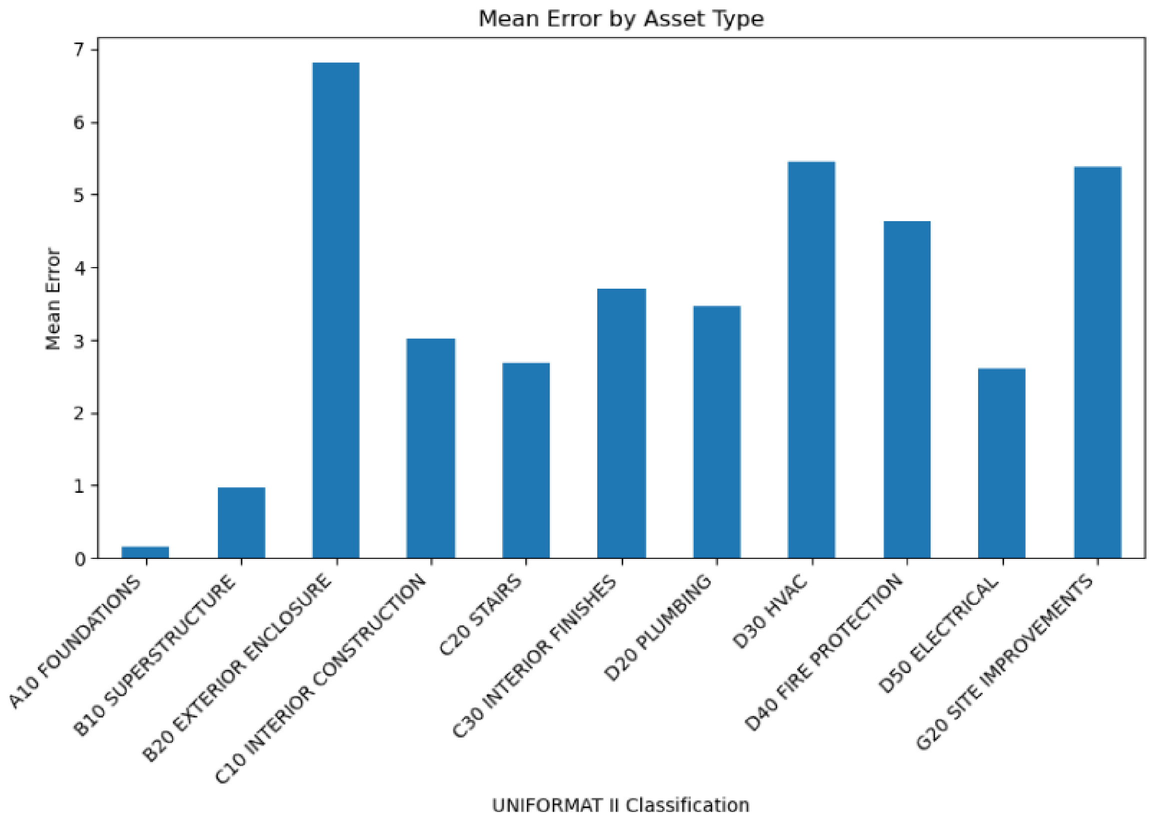

Figure 9. Mean Error of Inspections by UNIFORMAT II element classification.

Table 2. Summary Statistics of Inspection Errors by UNIFORMAT II element classification.

UNIFORMAT II Classification	Mean Error [-]	Standard Deviation [-]
A10 Foundations	0.15	0.25
B10 Superstructure	0.97	4.74
B20 Exterior Enclosure	6.82	17.73
C10 Interior Construction	3.01	9.67
C20 Stairs	2.69	9.82
C30 Interior Finishes	3.71	13.62
D20 Plumbing	3.46	11.22
D30 HVAC	5.45	15.66
D40 Fire Protection	4.63	12.86
D50 Electrical	2.60	9.24
G20 Site Improvements	5.37	16.09

Exploratory distribution plots of inspection frequency in **Figure 10** did not yield a distinct relationship between the frequency of inspections and the resulting condition index. This investigation was intended to highlight a trend between low assessed condition and increased inspection frequency but could not be validated in this dataset. Due to several outliers, a parameter was set to reduce this plot to only assets with less than 5 inspections per year.

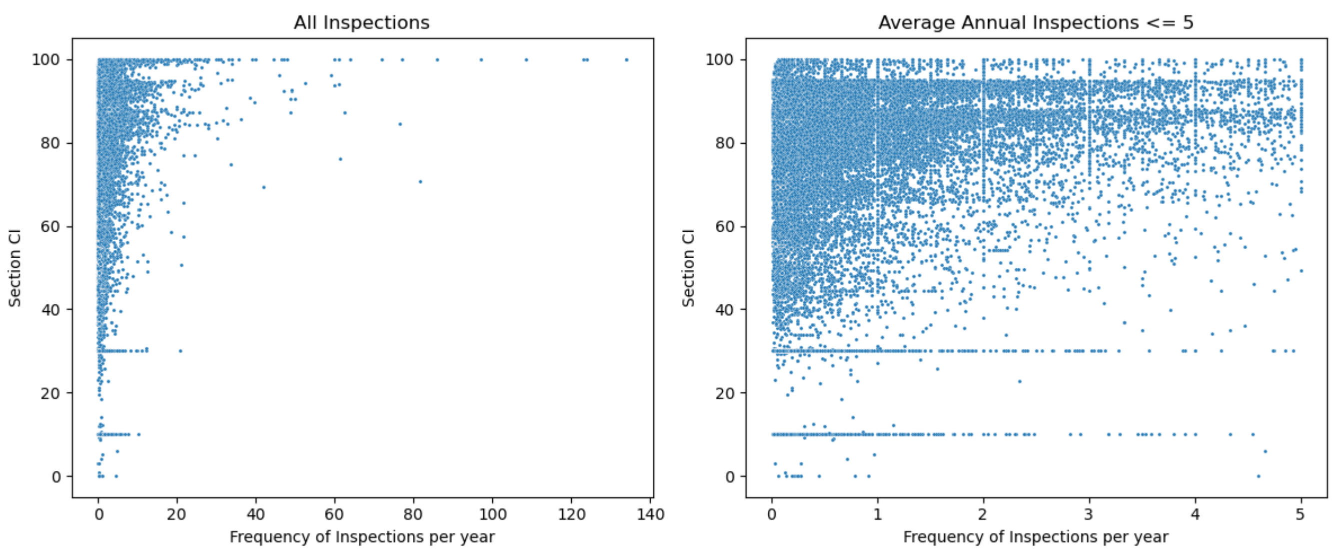

Figure 10. Distribution plot of average annual inspections per year versus asset condition index for: (a) all values of average annual inspections; (b) average annual inspections of 5 or less.

Figure 11 shows the CV distribution of error in inspections from a sample of unique inspectors. This figure shows a widespread with CV values ranging from 0.12 to 6.42, with over 93% of inspectors having a CV greater than 1. Ultimately this quantifies wide variability in error, and subsequent variability both between and among inspectors. CV values were calculated with respect to error with the standard deviation of error being 12.689. In total there were 1,122 unique inspectors in the dataset. Using the statistical formula for determining sample size in equation 2, a random sample of 582 was collected. A 95% confidence level with Margin of Error of 5% was used. From the sample of 582 inspectors, the CV was calculated for all inspections completed by that individual. To maintain anonymity, each of the 1,122 inspectors in the entire dataset was assigned a unique number, labeled as the Unique Inspector Index.

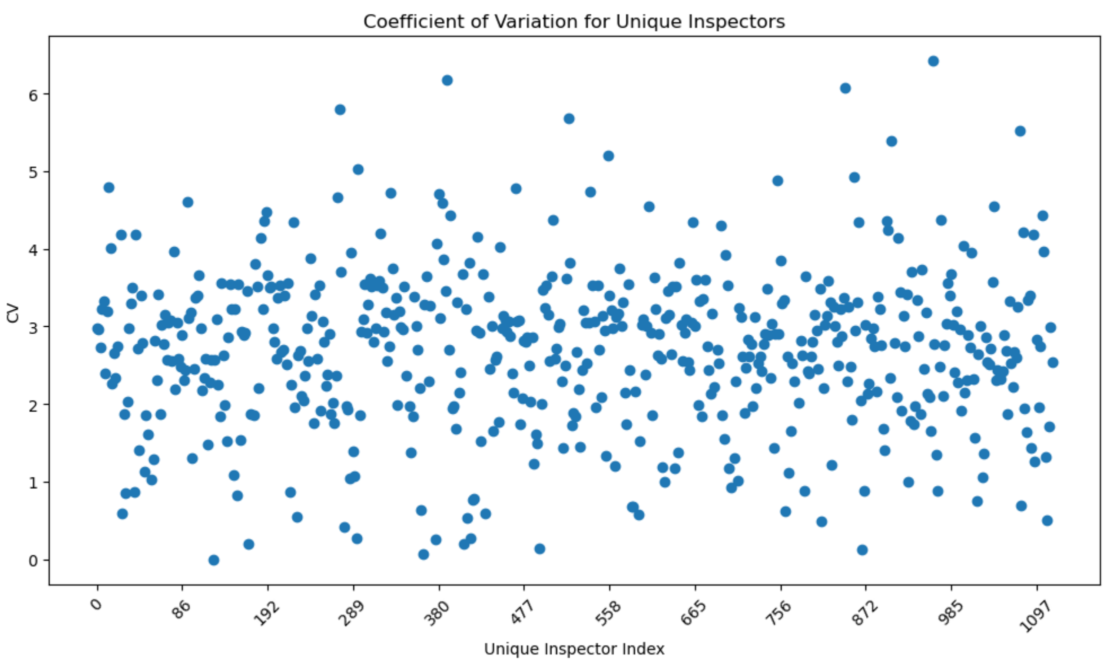

Figure 11. Coefficient of Variation values based on Error on a sample of 582 unique inspectors.

As a means of comparison between inspections by UNIFORMAT II element classification a Kruskal Wallis Test was conducted. The hypothesis test for the Kruskal Wallis test included the following null and alternate with a significance of α=0.05.

H_o: There is no significant difference between UNIFORMAT II Classification X's mean error and the mean sample error.

H_a: There is a significant difference between UNIFORMAT II Classification X's mean error and the mean sample error.

The result of this hypothesis test yielded a statistic of 15,156.32 with a p-value less than 0.0001 which rejected the null and indicated significant difference between mean error of the asset classifications. To further examine these results Tukey's HSD test was conducted. This compared each UNIFORMAT II element classification to another. The more specified null and alternate hypothesis used for this test, again with a significance of α=0.05, are shown below. To visualize the results of this test a spaghetti plot was used with A10 Foundations as the reference group, shown in **Figure 12**. Each point indicates the mean difference in error between the reference group with a 95 percent bound for confidence interval. If a confidence interval does not overlap between two asset groups, then the mean error can be considered significant between them.

H_o: There is no significant difference between UNIFORMAT II Classification X's mean error and UNIFORMAT II Classification Y's mean error.

H_a: There is a significant difference between UNIFORMAT II Classification X's mean error and UNIFORMAT II Classification Y's mean error.

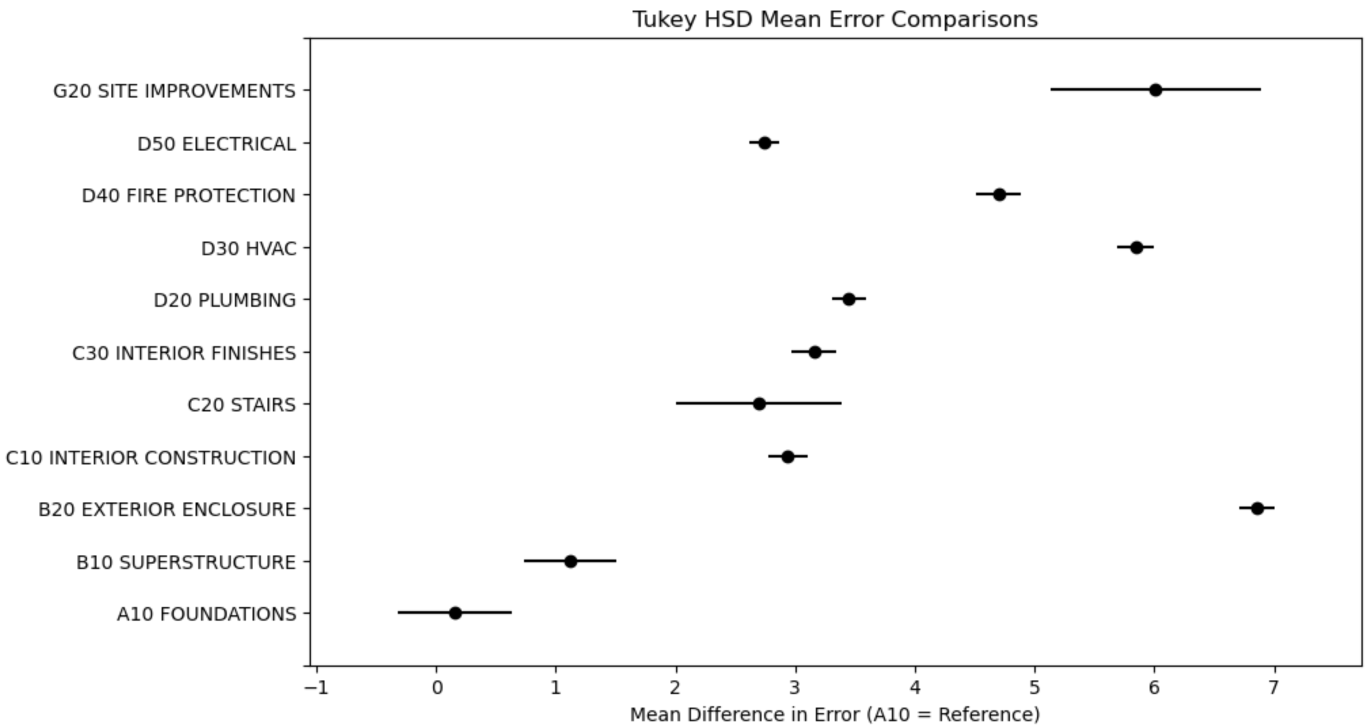

Figure 12. Spaghetti plot of Tukey HSD hypothesis tests results between UNIFORMAT II element classifications. A10 Foundations is the reference group.

Discussion

The analysis of the infrastructure inventory dataset revealed several important findings. Firstly, the distribution plots of Year Built, BCI, and MDI demonstrated non-normal distributions with localized peaks throughout. Spearman's Rho was employed to examine the relationships between Year Built and BCI, as well as Year Built and MDI. The results showed a weak, positive correlation between Year Built and BCI ($p=0.371$), which was statistically significant, and a negligible negative correlation between MDI and BCI ($p=-0.065$), also statistically significant. These findings were confirmed by visual analysis of heatmaps where Year Built exhibited a stronger positive correlation with BCI compared to MDI. This suggests that the age of infrastructure has a more substantial influence on its condition than its mission dependency. Intriguingly, MDI exhibited clusters of outlying regions, particularly around MDI values of 25 and close to 0. This suggests that certain facilities classified as non-critical for the Air Force's mission are still maintained in relatively good condition, possibly due to effective lifecycle management plans. These observations underscore the complexity of infrastructure management within the Air Force. Inspectors' evaluations of infrastructure conditions, when compared to the recorded condition indices in the BUILDER SMS system, revealed significant differences. Inspectors tended to rate fewer infrastructure assets as green and more as amber compared to the condition indices. This trend was visualized throughout all CONUS MAJCOMS indicating that location and mission did not have an impact on this discrepancy. The misalignment between inspector assessments and recorded conditions raises questions about the consistency and accuracy of the inspection process. Analyzing inspector data by UNIFORMAT II element classification allowed for the identification of variations in inspection ratings by infrastructure type. Of particular note is HVAC systems which showed a higher percentage of green inspections but a lower condition index score. This could be indicative of a false negative, or Type II error scenario. Conversely Superstructure infrastructure recorded a smaller percentage of green inspections but a higher mean condition index, indicative of a false positive, or Type I error. Hypothesis testing further revealed variations in inspector ratings across different MAJCOMs. The percentage of inspectors with significant differences from the population mean in terms of inspection rating varied considerably among MAJCOMs. This suggests that geographic or organizational factors may influence the consistency of inspectors. Additionally, all demonstrated inspection variability with an average of over 30% of inspectors differing from the population mean in green, amber, and red scoring categories.

The Coefficient of Variation analysis allowed for the assessment of individual inspector variability. From the sample size of 582 inspectors, only 40 (6.87%) had CV values less than one. This indicated that most inspectors varied drastically in error and in consistently matching inspections with expected scores. Additionally, this analysis provided a wide distribution of CV values with a range from 0.12 to 6.42. This distribution helps underscore the notion that there is variability in inspections between inspectors. The CV analysis can help identify inspectors who consistently provide more reliable assessments and those who exhibit greater variability. Understanding the factors contributing to this variability can inform training and quality assurance efforts. The Kruskal-Wallis test revealed significant differences in mean error across UNIFORMAT II element classifications, indicating that certain asset types are more challenging to assess accurately. The subsequent Tukey's HSD test further highlighted these differences, providing specific insights into which element classifications exhibited significant mean error discrepancies. The highest disagreement was in Foundations inspections which showed a significantly different mean error from all other asset classifications. Additionally, exterior enclosure, HVAC, and Fire Protection systems showed high levels of significant difference from most other elements. These findings can inform targeted improvement strategies for the inspection process.

Conclusions

The comprehensive analysis of inspection variability using Air Force infrastructure data yielded insights into the multifaceted realm of infrastructure management. Several key findings have emerged, shedding light on critical aspects of the Air Force's approach to infrastructure assessment and facilities maintenance.

Notably, the study highlighted variations in inspection ratings based on UNIFORMAT II element classifications, emphasizing certain elements, such as MEP systems, as areas of concern due to notable disparities between inspection ratings and condition indices. This calls for a deeper investigation into false negatives and false positives within these categories. The analysis of inspector variability, as assessed through CV calculations, indicated significant disparities among inspectors. Only a small fraction demonstrated consistently reliable assessments, highlighting the need to identify and address the contributing factors. Understanding the root causes of this variability will be instrumental in shaping comprehensive training programs and quality assurance efforts.

Additionally, the investigation uncovered a statistically significant yet weak positive correlation between infrastructure age, as represented by the Year Built metric, and the Building Condition Index (BCI). This observation suggests that aging infrastructure tends to exhibit slightly lower BCI scores. Furthermore, clusters of outlying regions comparing BCI and Mission Dependency Index (MDI) imply that facilities categorized as non-critical to the Air Force's mission still exhibit effective lifecycle management plans, resulting in their maintained good condition.

Despite these findings, it is essential to acknowledge the study's limitations. The analysis relies on existing data and assumes the accuracy of recorded condition indices produced by BUILDER SMS. Furthermore, while variability has been identified, further research is warranted to uncover the underlying causes and formulate strategies for its reduction. For future research endeavors, it is recommended to explore avenues for minimizing variability, both at the individual inspector level and across inspectors. This should involve the development of comprehensive training programs, revision of standardized inspection protocols, and identifying factors to improve quality control. Additionally, future investigations could focus on policy enhancements or inspection process updates for asset types that exhibited the highest levels of variability, such as MEP systems and Exterior Enclosure elements.

Journal Paper 2: Reducing Inspection Uncertainty in HVAC Systems: Strategies and Model Development

Reducing Inspection Uncertainty in HVAC Systems: Strategies and Model Development

Abstract

Within infrastructure asset management, this study investigates human inspections, with a particular focus on HVAC systems, using data sourced from the BUILDER Sustainment Management System (SMS) program. Despite the development of advanced technologies, manual inspections remain a dominant practice, prone to cognitive bias, variability, and uncertainty (Gordan et al. 2022). This research seeks to mitigate these challenges and reduce HVAC inspection variability. The implementation of predictive models, utilizing Linear, Random Forest, and Gradient Boosting Regression techniques, exhibits distinctive performances. Predicting Inspection Ratings based on input variables, including Age-Based Obsolescence, Remaining Design Life, Section Age, and Component, yielded reductions in Mean Squared Error (MSE) of 90.47% using Gradient Boosting Regression and reductions near 10% when compared to Expected Ratings. A Variability Rating prediction was also developed, incorporating Age-Based Obsolescence, Expected Rating, Section Age, Component, and Inspector as input variables. Both Random Forest and Gradient Boosting regression techniques displayed notably low MSE and high r^2 values, indicating a strong model fit. Specifically, Random Forest outperformed Gradient Boosting when exposed to cross validation, reflecting the potential in assessing the variability in inspection outcomes, a potential decision support tool for HVAC inspections. This study investigates the intricacies of HVAC inspection processes while also addressing variability and uncertainty in manual inspections. While predictive models for Inspection Ratings require further refinement to improve their representational accuracy, models for Variability Rating exhibit substantial

potential. Future research should focus on refining these models, enhancing their practical applications in HVAC maintenance, and exploring the incorporation of predictive models as decision support tools for infrastructure inspections.

Introduction

Human inspection is a pivotal component of infrastructure asset. Skilled inspectors are tasked with the evaluation of various civil infrastructure components, conducting assessments of their conditions, identifying maintenance needs, and determining necessary repairs. Despite the growing interest in advanced technologies and automated inspections, the use of human-in-the-loop procedures is still widely practiced. Humans are susceptible to cognitive bias, variability, or uncertainty when inspecting and finding solutions to reduce these factors is paramount (Gordan et al. 2022). Of particular interest is HVAC systems, which contribute 38 to 50 percent of total facility energy consumption (Cai et al. 2023; Tharanga et al. 2022; Yu et al. 2014). These systems are major contributors to building operations costs and can even reduce workplace productivity in periods of thermal discomfort (Cai et al. 2023; Lan et al. 2011). Maintaining these systems requires periodic inspection and maintenance, with energy efficiency shown to decrease by 20 to 40 percent when systems have thermal leakage (Aydin and Ozerdem 2006; Cai et al. 2023; Fisk et al. 2000). HVAC-specific inspections primarily require human involvement or analysis and are difficult, time-consuming, and prone to human error (Cai et al. 2023; Torabi et al. 2021). Finally, previous research has shown HVAC inspections to have one of the highest degrees of variability among inspectors when compared to similar infrastructure systems (Pratt 2023).

In this context, the purpose of this study was to explore issues with human inspection of HVAC systems and develop a model to aid in variability reduction. By reviewing previous literature on HVAC predictive modeling and inspection uncertainty, a combined model can be developed to address areas of concern and a wholistic approach to improvement can be recommended.

Background

Current State of HVAC Inspections

Beginning with construction, research shows that initial installation of HVAC systems is the second most prominent concern for defects when inspected (Tekin et al. 2023). After installation of an HVAC system, operation of this infrastructure brings additional concerns with respect to inspections. At its current stage, HVAC inspection relies heavily on human involvement, requiring inspectors with qualified domain knowledge (Cai et al. 2023; NADCA 2021). Specific examples include visually inspecting for leakages in air ducts through smoke testing, and pressure testing ducts through manual pressure gauge monitoring (Kute et al. 2015; Mathaudhu 1997). As a result, these inspections are known to be reactive, time-consuming, and costly (Cai et al. 2023; Hegazy et al. 2010; Lan et al. 2011). At the policy level, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) and the National Air Duct Cleaners Association (NADCA) both have standards that recommend quarterly, semiannual, or annual inspections (ASHRAE 2018; NADCA 2021). Additionally, workplace productivity has shown reduction when thermal discomfort exceeds a Predicted Mean Vote (PMV) range from -0.5 to 0.5 (ASHRAE 2018; Lan et al. 2011). This demonstrates that an improperly maintained and inspected HVAC system can have secondary and tertiary effects, beyond the operation of the facility itself. With respect to inspections, issues with HVAC systems are usually only detected when thermal discomfort is met by occupants, making the process more reactive than proactive (Cai et al. 2023; Lan et al. 2011). Given that manual inspection of HVAC systems is prone to subjectivity and inspector bias, efforts have been made to use technology to improve the process. Using applications common in sewer pipeline inspection, robots with camera systems have been developed to inspect HVAC ducts, but this process still requires analysis by domain experts (Wang and Su 2014). Efforts have been made to develop autonomous robots and Unmanned Aerial Vehicles (UAVs) for inspection of HVAC systems. At its current state, UAVs lack the ability to focus specifically on HVAC system parameters, and autonomous robots cannot intrusively inspect ducts systems with the system in operation, highlighting technological drawbacks on these inspection techniques (Cai et al. 2023). Additionally, installation of sensors for continuous monitoring has shown promise in detecting defects in air handling units and other key HVAC components (Lei et al. 2021; Zhang et al. 2021). While this technology is good for newly constructed buildings, it is difficult and costly to retrofit into aged facilities and infrastructure (Ricketts 2016). This field of research supports advancement in technology for HVAC system inspection, but the reliance on skilled human involvement is still necessitated.

Factors That Affect HVAC Inspection Variability

The dominant means of human evaluation of HVAC systems is visual inspection. While there are techniques for measuring and quantifying parameters such as thermal leakage and pressure, visual inspection allows for identifying defects, cracks, component wear, and various other issues within the system. Broadly, human-centered visual inspection is prone to cognitive bias and decision bias which dictates variability in findings due to differences in judgment and evaluation from inspectors

(Campbell et al. 2021; Gordan et al. 2022). The manufacturing process for automotive engines uses similar visual inspection techniques to that of HVAC systems. This process of parts inspection is prone to human error due to factors such as lighting, experience, complexity of the task, and education on the inspection procedure (Dubey et al. 2023; Stallard (Voelker) et al. 2018). Similarly, techniques for bridge defect inspection also align similarly with HVAC inspections. In this procedure, inspectors identify, analyze, and prioritize defects (cracks) in bridge components to implement a repair and rehabilitation strategy (Liu et al. 2023). In this regard, inspector experience has shown more consistent inspection results with reduced variability (Ball et al. 2017; Woodcock 2014). Specific factors effecting inspection quality in this field are attributed to work experience, training, environmental and site conditions, and demonstrating an inspection task prior to completing an assessment (Liu et al. 2023).

Particularly with HVAC systems, inspections are conducted for fault detection and diagnosis (FDD). These faults can be classified as hard (i.e.: pressure drop, decrease in motor efficiency, air leakage, and fouling in air handling units) or soft (i.e.: tuning errors, poor installation, non-optimal commissioning) (Torabi et al. 2021; Yu et al. 2014). Both hard and soft faults are attributed with increased energy consumption, energy waste, and equipment failure making the need for accurate FDD important (Torabi et al. 2021; Yu et al. 2014). Research by Torabi et al. identified human factors impacting HVAC equipment life and energy consumption, being non-optimal commissioning, poorly developed Operations and Maintenance manuals, infrequent maintenance intervals, and overdue preventative maintenance (2021). Summarily, experience, education on evaluation procedure, and inspection frequency are all key factors contributing to HVAC inspection variability.

Strategies To Improve HVAC Condition Determination

Review of literature on HVAC system condition determination primarily focused on modeling, soft computing, and machine learning (ML) capabilities. In most instances, inspection of HVAC systems is used to drive a condition assessment that can be used as a decision support tool for prioritization of maintenance, repair, and replacement actions. One example is the development of a probabilistic model to reduce Type I and Type II errors in crack detection. This model leveraged subject matter experts for probability determinations of factors including lighting, experience, and training, and incorporated a ML algorithm to achieve an overall accuracy of 89% (Dubey et al. 2023). In this instance training and experience were two of the top three factors contributing to inspection variability.

When determining the condition of an HVAC system one approach demonstrated success using soft computing. Fuzzy sets theory was applied to incorporate uncertainty when classifying HVAC components condition on a variety of factors. The fuzzy sets theory is used to translate uncertain classifications such as “poor”, “fair”, and “good” into quantitative numerical observations (Chan et al. 2009). With this approach, Besiktepe et al. used data on Mean Time Between Failure, Age-Based Obsolescence, Facility Condition Index, Occupant Feedback, and Preventive Maintenance Cycle and applied a fuzzy set classification to develop membership functions and ultimately determine the condition of a 38 year old Chiller on a scale from 1 to 5 (2021).

Another strategy for determining HVAC system condition involves the use of modeling with factors directly related to the equipment itself. Through collection of data on frequency of breakdowns, equipment downtime, and cost of repair an asset condition can be predicted and applied to an established decision-making framework to determine the necessary maintenance actions (Ma et al. 2020; Rastegari and Mobin 2016). This claim is supported by an expert panel survey that indicated evaluation of physical condition and frequency of failures as 34.2% and 25.7% important in determining asset condition of HVAC systems, respectively (Matos et al. 2023). Similarly the influence of HVAC equipment runtime was investigated and used to develop deterioration curves comparing both a probabilistic and nonlinear optimization model (Tharanga et al. 2022). In this instance, components such as air handling units, chillers, and exhaust fans, were classified based on their weekly runtime and used to develop deterioration curves. The results of this study showed HVAC components anywhere from 6 to 16 percent increases in deterioration rates when compared to curves that did not classify based on runtime. Ultimately this review showed substantial literature in determining the condition of HVAC systems, but all lacked the incorporation of parameters concerning uncertainty and variability in inspections.

Research Questions

The objective of this research was to identify and evaluate strategies for reducing human inspection uncertainty in HVAC systems and to investigate the feasibility of developing a model to achieve this reduction. The primary research questions to be answered are listed below. Based on these research questions a review of current literature was conducted to determine the current state of HVAC inspection, evaluate parameters that affect HVAC inspection variability, and identify strategies to improve HVAC condition determinations.

1. What strategies exist to reduce human inspection uncertainty in HVAC systems?
2. Can a model be developed to reduce inspection uncertainty in HVAC systems?

Methods

Data Set Selection

Based on the review of literature there is a gap in research that aims to reduce HVAC inspection variability. Therefore, a need was identified to quantitatively answer research question two. The primary means to answer the research questions was through development of a predictive model that would incorporate parameters specifically tied to inspection variability. This involved developing a model to reduce inspector variability by predicting an asset condition for decision support. To develop, test, and validate this process a case study using data from the BUILDER Sustainment Management System (SMS) was created. BUILDER SMS was created in 2009 and is the primary facility asset management tool used by the Department of Defense (DoD) ("BUILDER Sustainment Management System" 2012). This system inputs data from visual inspections of DoD personnel and uses that to generate condition assessments of all infrastructure assets. From there, BUILDER SMS can be used as a tool to make decisions on maintenance actions, work prioritization, and capital project investments. Additionally, all infrastructure assets contained within BUILDER SMS are classified using the ASTM UNIFORMAT II building element classification scheme, making this easy to categorize into HVAC components. The process for inspecting HVAC systems in the DoD involves skilled inspectors visually assessing individual components of the system and assigning a rating of green, amber, or red with a plus, neutral, or minus additive and inputting into the BUILDER SMS system (USACE 2017). Definitions of different ratings based on condition, operational performance, maintenance requirements, and distress are shown in **Table 3**, below. One important note is that components that are not visible are only inventoried, and their condition assessment is calculated purely using deterioration curves generated by the BUILDER SMS program.

Table 3. Inspection Rating Criteria for HVAC Systems in BUILDER SMS. Adapted from USACE 2017.

Rating	Condition	Operational Performance	Maintenance Required	Distress
Green +	No deterioration	Fully operational. Normal PM required.	Distresses present have no impact to operations. No maintenance outside of normal operations.	
Green	Slight deterioration	Fully operational. Normal PM required.		
Green -	Noticeable deterioration	Fully operational. Normal PM required.		
Amber +	Minor deterioration	Operation slightly affected. Repair required.	Distresses present impact operations and repair is needed to return to proper operations. Maintenance outside of normal operations required.	Minor/Mild
Amber	Moderate deterioration	Operation moderately affected. Repair required.		Moderate
Amber -	Considerable deterioration	Operation considerably affected. Repair required.		Major/Considerable
Red +	Significant deterioration	Operation significantly affected. Repair required.	Distress present impact operations and replacement is needed. Maintenance will not fix the problem at hand.	Significant/Extensive
Red	Severe deterioration	Operation severely affected. Repair required.		Severe
Red -	Complete deterioration	No longer operational. Replacement required.		Complete

Data Collection

Data collection involved generating several pre-existing reports from BUILDER SMS over a collection period from February 2023 to May 2023. An Inspection Matrix report was generated on 16 active-duty Contiguous United States (CONUS) Air Force Bases to provide data on all inspections that had been completed and reported in BUILDER SMS. Due to the size of this dataset, a random sample of 16 installations was used. This sample ensured that every Major Command (MAJCOM) had at least 2 installations in the sample size to ensure differences between locations and mission sets were captured. Information on the data provided in the Inspection

Matrix report can be found in Appendix A. All reports were combined into a working data frame that was used for data analysis in a Juptyer Notebook with Python coding.

Data Preprocessing

Following collection of the dataset preprocessing was required to create a compatible data frame for model development. The following actions to preprocess the dataset are detailed in step-by-step fashion below.

1. **Isolate to HVAC Systems.** Inspection data was isolated to only components classified as D30 HVAC in UNIFORMAT II criteria to reflect infrastructure attributed to HVAC systems.
2. **Remove Outlier Ratings.** Any inspection rows with Expected Rating scores that were missing or had a value of zero were removed as outliers. This left a dataset of 91,691 unique inspections covering a time period from 1992 to 2023.
3. **Remove Distress Inspections.** Inspections that were classified as distress rather than direct ratings were removed as outliers. These distress ratings are atypical and follow a different inspection process (USACE 2017). This could contribute to variability in results not attributed to an inspector itself and therefore was removed, resulting in a dataset of 90,547 unique inspections covering a time period from 1992 to 2023.
4. **Remove Unknown Inspectors.** For this action, any inspection where the inspector was not listed was labeled as “Unknown”, “Inspector”, or “Inspection Supervisor”. These inspections, totaling 969 of the dataset, were removed because characteristics on inspector variability could not be applied.

5. **Calculate Squared Error Feature.** An additional feature for squared error was calculated and added using the formula given in Equation 1, below. This is a standard calculation used for analysis of model fit, calculation of Mean Squared Error (MSE), and can also be used to assess variability reduction in inspections. Values for squared error help illustrate the magnitude of difference between two variables.

$$\text{Squared Error} = (\text{Expected Rating} - \text{Inspection Rating})^2 \tag{1}$$

6. **Calculate Scaled Variability Rating.** The creation of this feature was used to generate an easily understood scale that would classify the degree of variability in a given inspection. This was done by normalizing the Squared Error feature and then scaling to a 1 to 10 scale. To transform the data the ladder of powers concept was adapted using techniques such as square root, logarithmic, and reciprocal, with cube root achieving the closest degree of normality (Velleman and Hoaglin 1981). The calculation for this is shown in Equation 2, below. The transformed data was then normalized with the Min-Max technique using Equation 3 and scaled from 1 to 10 using Equation 4 to create a Variability Rating (Patro and Sahu 2015).

$$\text{Transformed Squared Error (TSE)} = \sqrt[3]{\text{Squared Error}} \tag{2}$$

$$\text{Normalized Squared Error (NSE)} = \frac{\text{TSE} - \text{Min TSE}}{\text{Max TSE} - \text{Min TSE}} \tag{3}$$

$$\text{Variability Rating} = \text{min NSE} + (\text{max Variability Rating - min Variability Rating}) * \text{NSE} \tag{4}$$

7. **Calculate Percent Error Feature.** This is another common metric used for analyzing and understanding variability between inspections and inspectors. This metric for percent error gives a ratio of how different an inspection rating is from the expected rating calculated by BUILDER SMS. The calculation for this feature is shown in Equation 5, below. Values for percent error typically range from 0 to 1. The non-absolute value percent error was also calculated to help observe undervalued and overvalued inspections using Equation 6, below. In this case, negative values indicate an undervalue by BUILDER SMS where inspection shows a higher condition index than expected. Conversely, positive values indicate an overvalue where inspection shows a lower condition index than expected.

$$\text{Percent Error} = \left|\frac{\text{Expected Rating-Inspection Rating}}{\text{Expected Rating}}\right| \qquad (5)$$

$$\text{Non} - \text{AB Percent Error} = \frac{\text{Expected Rating-Inspection Rating}}{\text{Expected Rating}} \qquad (6)$$

8. **Calculate Age-Based Obsolescence.** This variable was identified in work by Grussing and Besiktepe et.al as a ratio index that can be used to illustrate equipment functionality beyond its expected useful life (2021; 2014). Adapting an Age-Based Obsolescence variable to this data set used Equation 7, below. Values for Age-Based Obsolescence ranged from 0 to 1 with values closer to zero showing infrastructure in good condition relative to its service life and decreasing in time-based condition as it approaches 1.

$$\text{Age Based Obsolescene} = \frac{\text{Section Age}}{(\text{Current Year} - \text{Section Install Date}) + \text{Remaining Service Life}} \qquad (7)$$

9. **Convert Categorical Variables.** Within the QC-6 report categorical variables exist that needed conversion for data preprocessing. For this process one-hot encoding was used. This process converts *n* options for a categorical variable and creates an *n*-dimension matrix with a 1 used to designate matching an option and a 0 to indicate not matching other options (Seger 2018). In this instance, Component and Inspector variables were identified as categorical variables and one-hot encoded using the Pandas Python library.

Data Exploration

Data exploration involved visualization and statistical analysis of significant variables in the dataset for any trends, patterns, or outliers. Visualization included generating histograms of numerical variables to identify any obvious skew, normality, or outliers. Additionally, scatterplots of each individual variable were generated against the Expected Rating (the dependent variable of interest) to identify any potential trends that could be used for regression. Statistical analysis involved calculating summary statistics (mean, median, maximum, minimum, standard deviation) for each continuous variable. Correlation coefficients were also calculated between variables to statistically identify any potential relationships. For continuous variables Pearson's R or Spearman's Rho Correlation Coefficient can be used, while nominal categorical variables (Component and Inspector) require the use of a Kramer's V Correlation Coefficient (Akoglu 2018; Artusi et al. 2002). A Summary of the variables, definition, and data classification is shown in **Table 4**, below.

Table 4. Summary of variables analyzed in Data Exploration.

Variable	Definition	Data Type
Age Based Obsolescence	Ratio of component age to expected service life	Continuous
Component	UNIFORMAT II classification of HVAC component	Categorical
Expected Rating	The expected condition index for an inspection generated by BUILDER SMS	Continuous
Inspection Rating	The condition index for an asset based on the inspection results	Continuous
Inspector	The unique value given to each inspector for differentiation	Categorical
Non-AB Percentage Error	The percentage error calculated without absolute value	Continuous
Paint CI	The condition index of the paint specifically associated with a component	Continuous
Percentage Error	The percentage error between inspection and expected rating	Continuous
RSL (Years)	The remaining service life calculated by BUILDER SMS in years	Continuous
Section Age	The age of the entire HVAC system attributed to a specific component	Continuous
Section CI	The condition index of the entire HVAC system attributed to a specific component	Continuous
Section RDL	The remaining design life of the entire HVAC system attributed to a specific component	Continuous
Squared Error	The calculated squared error between inspection and expected rating	Continuous
Variability Rating	The index of variability from inspection scaled from 1 to 10	Continuous

The final portion of data exploration involved visual comparison of categorical to continuous variables. Inspectors were assigned a unique inspector number to maintain anonymity and the mean squared error of their respective inspections was plotted to visualize the distribution of variability among inspectors. Further, violin plots of variables such as expected rating, percentage error, and squared error were created by the component categorical data to see if there were differences in inspection variability among different HVAC systems.

Data Splitting

For training and testing of models the dataset was split into training and test components. Common splits include 60-40, 70-30, 80-20, and 90-10 for training and testing, respectively. An 80-20 split was used, which has been shown to have the highest accuracy in similar data intensive fields in medical research (Bhanot et al. 2018). This meant that the pre-processed dataset of HVAC inspection data would be randomly split, with 80 percent of the data being used to train a model and 20 percent being used to test the accuracy of the model.

Model Selection

To answer research question two, two distinct model methodologies were developed in an attempt to reduce HVAC inspection variability. The first model was aimed at predicting the Inspection Rating using different variables from the dataset. If the model could establish Inspection Ratings that were more accurate when compared to the Expected Rating, then these values could be used as decision support tool for inspectors. This could be implemented and notify inspectors of the predicted condition of an asset and be used as a reference to ensure abnormal inspection ratings received a level of due diligence before being recorded. Based on the dataset and results from correlation of variables Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor models were selected. Linear regression is one of the simplest model techniques and is based on a linear relationship between a dependent variable and one or more independent variables. Random Forest regression is a strong option for complex relationships, non-linearity, and a combination of numerical and categorical data (Kadam et al. 2020). In a similar manner to Random

Forest, Gradient Boosting Regression combines decision trees to promote high accuracy while also being able to handle different data types and relationships (Huang et al. 2019).

The second model was aimed at predicting infrastructure that may be susceptible to high inspection variability. In a similar manner to the first set of models, this could be implemented as a decision support tool for inspectors. The goal of this model would be to take characteristics of the HVAC infrastructure such as Section Age, Age-Based Obsolescence, and Expected Rating and combine this with the Inspector to predict a Variability Rating. Using this information, a model could be developed to alert if the potential for high inspection variability was expected and could then be messaged as a form of due diligence during inspection. Similar to the first model, this second model would utilize regression for prediction and given the same dataset a Linear, Random Forest, and Gradient Boosting technique were all tested.

Model Evaluation

To assess the performance of the models, different quantitative and qualitative measures were undertaken for validation. Using the predicted results from the models, comparisons were made between the actual results from the test sets using training accuracy, testing accuracy, MSE, r-squared, and reduction in MSE. These values are all standard quantitative measures of measuring model performance and fit. Qualitatively, distributions of the predicted values and expected values were plotted to visually confirm performance and model fit.

Results

Data Analysis

Initial analysis of the dataset visualizing the distributions of all potential variables is shown in **Figure 13**, below. Histogram distributions showed key variables such as Section RDL and Section Age exhibiting normality with other variables such as RSL, Section CI, Age-Based Obsolescence, and Percentage Error illustrating skew. Other variables like Paint CI and Inspection Rating did not show any distinct patterns or distributions but had regions of high and low frequency.

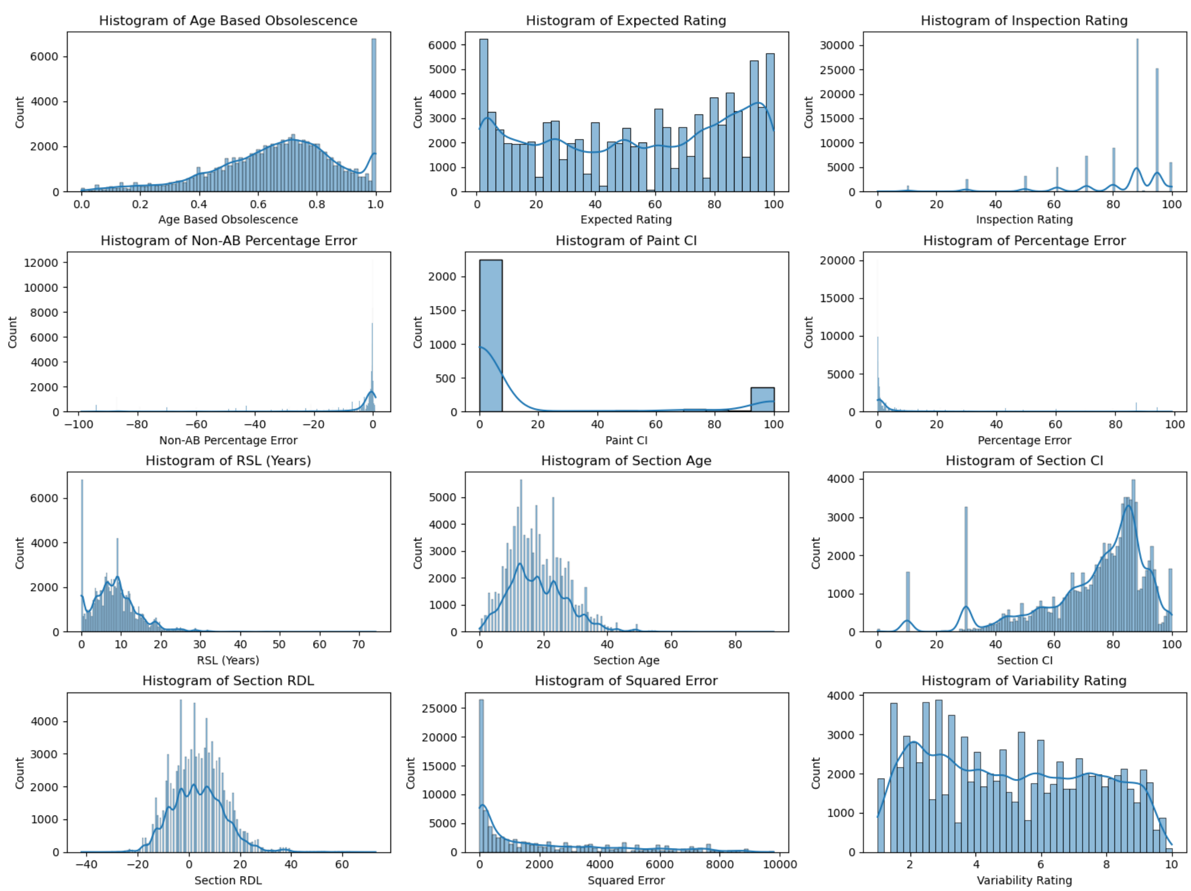

Figure 13. Histogram plots of key variables of the HVAC dataset.

Results of visual analysis between variables show potential relationships with Age-Based Obsolescence, RSL, RDL, Section CI, and Section Age when compared to Expected Rating, as illustrated in **Figure 14**. This analysis involved plotting potential variables against the key dependent value, Expected Rating. As the major variable used for comparison of error and variability, plotting this against other features could identify any apparent relationships that could be included when developing a regression model.

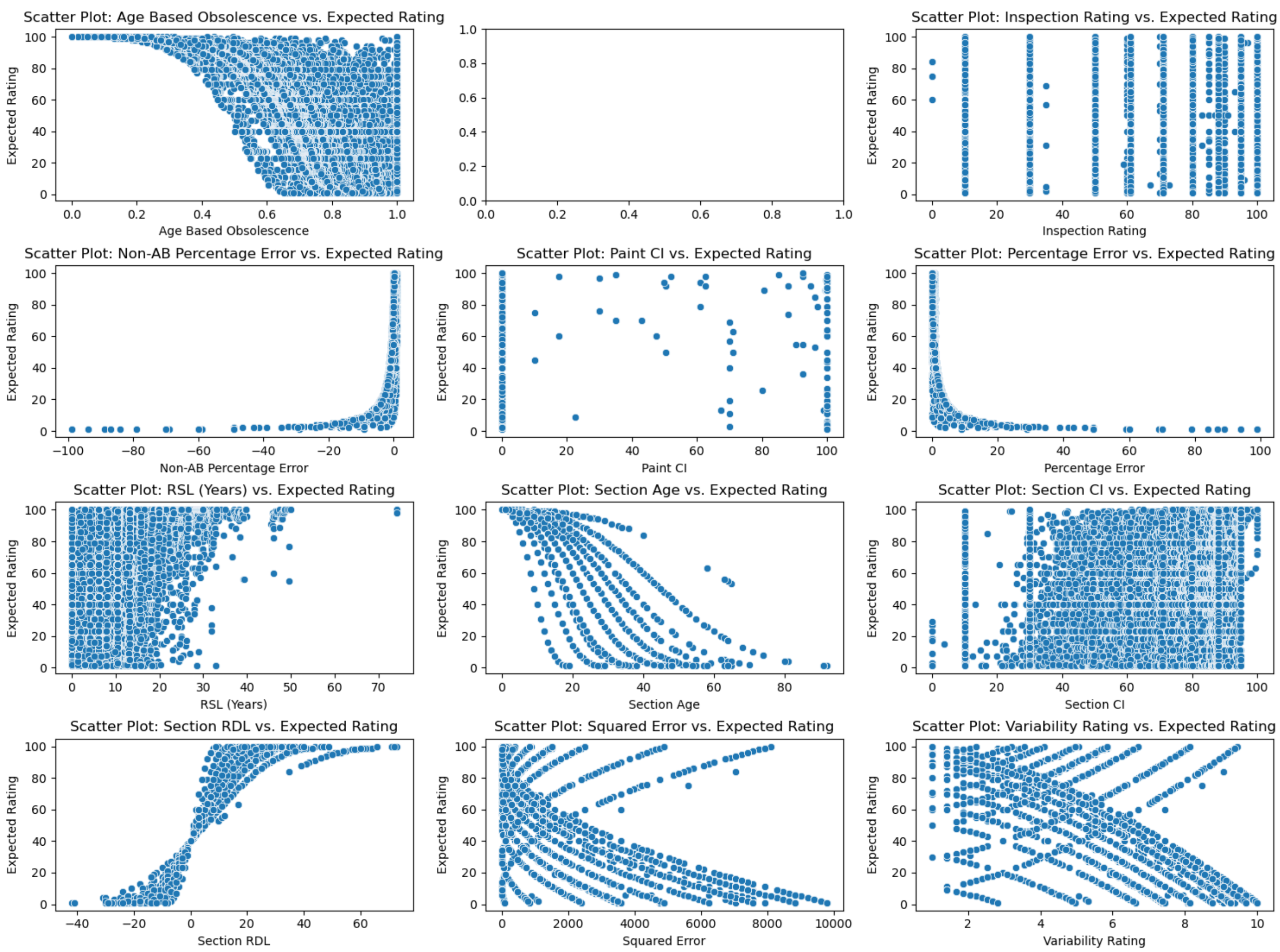

Figure 14. Scatterplots of key variables versus Expected Rating.

Results of summary statistics on key variables in **Table 5** show wide distributions of values for Expected Rating, Paint CI, Percentage Error, RSL, Section RDL, and Squared Error with standard deviation being as high or higher in magnitude than mean. This quantitative analysis of the dataset acted as a secondary confirmation of relationships seen in visual analysis. The primary summary statistics of concern were mean, standard deviation, median, maximum, and minimum.

Table 5. Summary Statistics of Key Variables in the HVAC dataset.

	Mean	Standard Deviation	Median	Max	Min
Age Based Obsolescence	0.670	0.209	0.690	1.00	0.00
Expected Rating	53.633	32.450	56.000	100.00	1.00
Inspection Rating	83.185	17.019	88.000	100.00	0.00
Non-AB Percentage Error	-5.382	15.671	-0.467	1.00	-99.00
Paint CI	16.117	35.329	0.000	100.00	0.00
Percentage Error	5.441	15.650	0.500	99.00	0.00
RSL (Years)	8.603	5.925	8.152	74.18	0.00
Section Age	18.234	9.141	17.000	92.00	0.00
Section CI	73.835	18.680	79.258	100.00	0.00
Section RDL	3.762	10.403	3.000	73.00	-42.00
Squared Error	1929.653	2355.099	784.000	9801.00	0.00
Variability Rating	5.052	2.480	4.878	10.00	1.00

The results of a Spearman's Rho correlation analysis in **Table 6** indicate that Expected Rating has potential correlations with Age-Based Obsolescence, Percentage Error, Section Age, Section RDL, Squared Error, and Variability Rating. Additionally, Age-Based Obsolescence had a strong correlation with many variables, most likely because it is derived from the Section Age, and RSL variables. The Spearman's Rho does not look at a linear correlation between variables, but rather a monotonic relationship (Akoglu 2018). Values highlighted in green show a moderate or very strong positive relationship, while values in orange or red show a moderate or very strong negative relationship. Since the Variability Rating is a normalized and scaled Squared Error term, the Spearman's Rho values are the same as those of Squared Error.

Table 6. Spearman's Rho Values for Correlation of Continuous Variables.

	Age Based Obsolescence	Expected Rating	Inspection Rating	Non-AB Percentage Error	Paint CI	Percentage Error	RSL	Section Age	Section CI	Section RDL
Expected Rating	-0.807									
Inspection Rating	-0.527	0.344								
Non-AB Percentage Error	-0.640	0.923	0.064							
Paint CI	0.157	-0.116	-0.117	-0.081						
Percentage Error	0.715	-0.931	-0.261	-0.913	0.088					
RSL	-0.833	0.512	0.461	0.358	-0.096	-0.440				
Section Age	0.669	-0.822	-0.368	-0.733	0.136	0.742	-0.222			
Section CI	-0.801	0.406	0.548	0.231	-0.132	-0.320	0.855	-0.337		
Section RDL	-0.786	0.978	0.330	0.910	-0.110	-0.918	0.540	-0.765	0.394	
Squared Error	0.621	-0.844	-0.139	-0.854	0.035	0.965	-0.356	0.658	-0.228	-0.831
Variability Rating	0.621	-0.844	-0.139	-0.854	0.035	0.965	-0.356	0.658	-0.228	-0.831

For categorical variables, the Cramer's V between Component and Inspector was found to be 0.261 which correlates to a very strong relationship (Akoglu 2018). This calculation allows for correlation between two categorical variables. Additionally, visual analysis involved plotting the MSE by Inspector, shown in **Figure 15**, below. This plot randomly selected 44 of the 440 unique inspectors in the dataset and plotted the mean squared error for all of their respective inspections. Additionally included was a bar showing one standard deviation. From this plot in **Figure 15** there is significant visual variability in error between inspectors, and there is also variability in error between inspections conducted by each inspector. To maintain anonymity each inspector was assigned a unique Inspector Number.

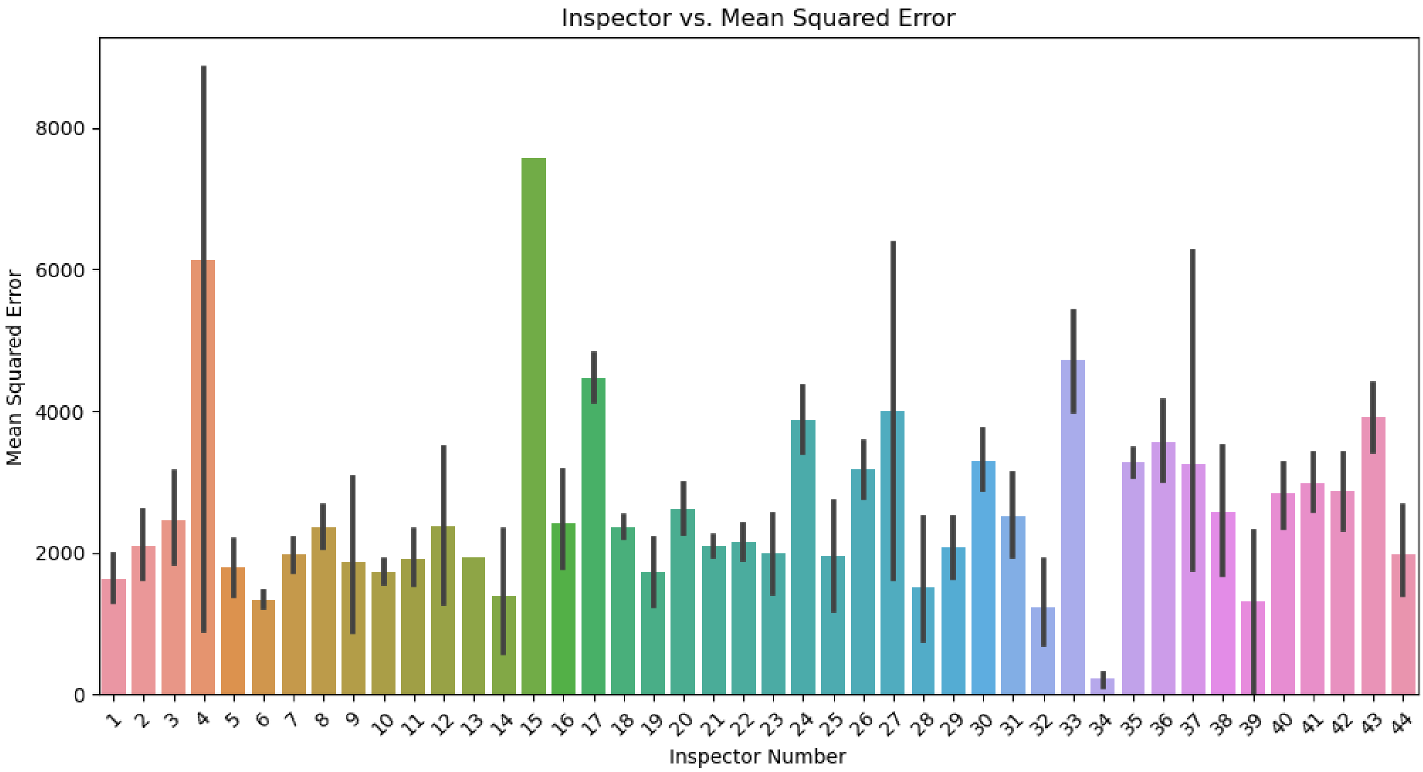

Figure 15. Mean Squared Error by Inspector with One Standard Deviation shown.

The violin plots of the component variable in **Figure 16** indicate that the component has little difference between percentage error, but that Age-Based Obsolescence and Expected Rating did have noticeable distinctions between different HVAC assets. These violin plots were used to visualize the Component variable and were plotted against key variables such as Age-Based Obsolescence, Expected Rating, Percentage Error, and Squared Error.

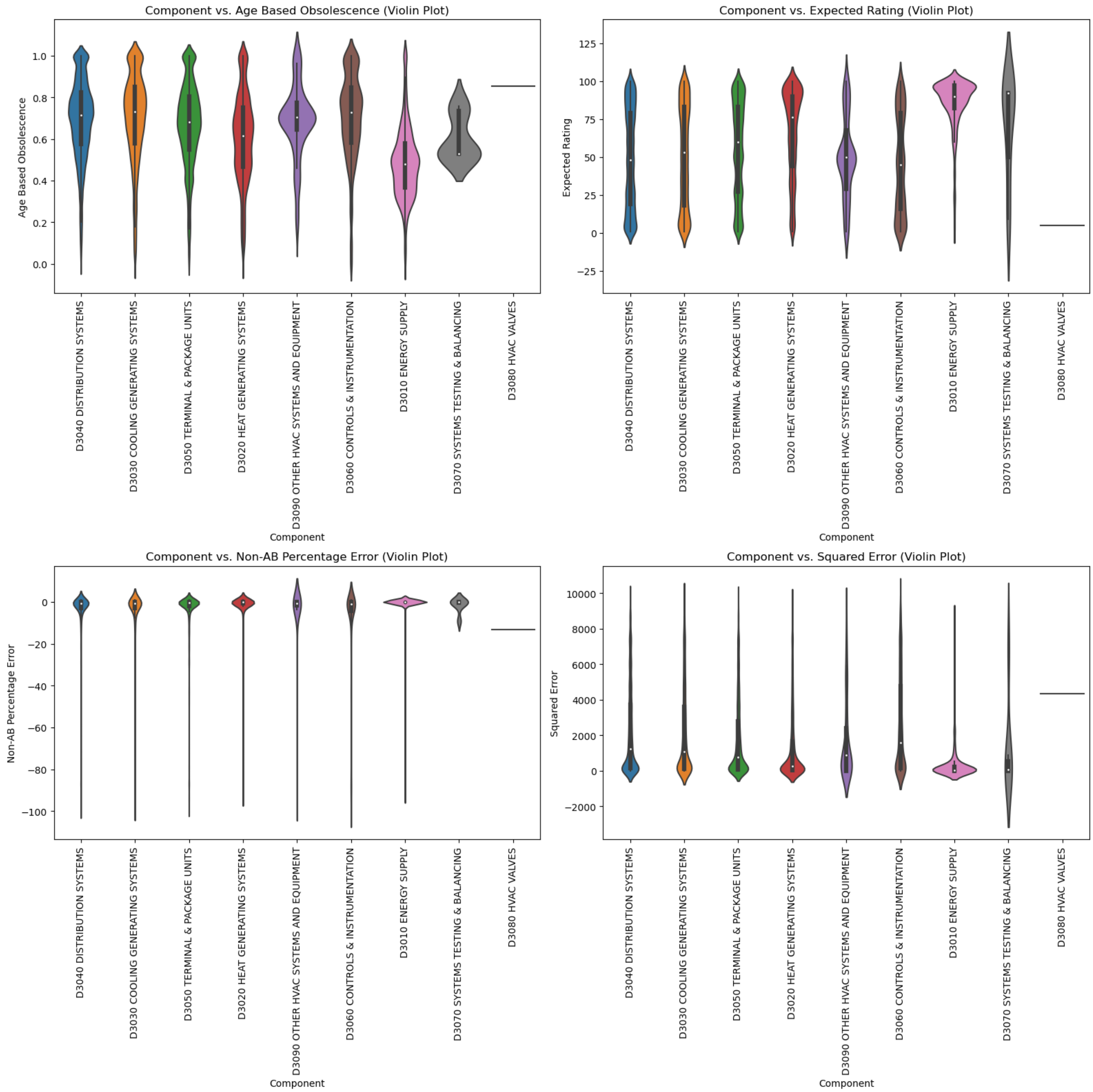

Figure 16. Violin Plots of HVAC components versus several continuous variables.

Data Modeling

The first set of models developed were predictive regression models aimed at predicting an Inspection Rating based on several input parameters. Ultimately, Age-Based Obsolescence, Section RDL, Section Age, and Component were selected as input variables. The resulting models were measured for performance and fit by calculating values for training accuracy, testing accuracy, MSE, r-squared, and reduction in MSE from the actual dataset. Results in **Table 7** summarize the performance values for each of the three tested models in predicting Expected Rating and comparing to the actual Expected Rating. The results from **Table 7** show significant reduction in MSE from the initial dataset of 1,929 with Gradient Boosting and Random Forest being nearly identical and linear regression slightly less effective. All r-squared values were relatively low showing that model fitting of the data was low in correlation. This was further supported with the low training and testing accuracies calculated for each respective model.

Table 7. Results of predictive model on Inspection Rating when compared to Inspection Rating.

Regression Technique	Training Accuracy	Testing Accuracy	MSE	r^2	MSE Reduction
Linear Regression	0.228	0.227	217.410	0.227	0.887
Random Forest	0.558	0.331	188.183	0.331	0.902
Gradient Boosting	0.365	0.347	183.692	0.347	0.905

One important note to these performance measures is that they measure how well a predicted Inspection Rating matches with the actual Inspection Rating. Since there is known variability between the Inspection Rating and Expected Rating, another measure of fit would be calculating these performance measures by comparing the predicted Inspection Rating to the Actual Expected Rating. This was conducted with results shown in **Table 8**, below. These performance measures show much smaller reduction in MSE with linear regression being the best overall model. While there is still reduction in overall error from these models, the r-squared values show little to no correlation making them unreliable for predicting Inspection Ratings that would better represent the anticipated Expected Rating. Since these models were developed to predict Inspection Rating, rather than Expected Rating, training and testing accuracies were not calculated.

Table 8. Results of predictive model on Inspection Rating when compared to Expected Rating.

Regression Technique	MSE	r^2	MSE Reduction
Linear	1709.075	-0.638	0.114
Random Forest	1809.665	-0.735	0.062
Gradient Boosting	1750.464	-0.678	0.093

To improve the performance of these regression models the Component categorical variable was removed and models were generated on each individually. This was used to limit generalization of the training set and serve as a preliminary classification procedure. The calculated performance measures of these models are shown in **Table 9**. Models for D3070 Systems Testing and Balancing, and D3080 HVAC Values were not considered due to a lack of data in the data set. When comparing to the original model, predictive models were significantly better with most values for accuracy, MSE, r-squared, and MSE reduction improving. In this instance, Gradient Boosting Regression models were generally the best with all except D3030 Cooling Generating Systems improving in model fit and accuracy.

Table 9. Results of component predictive models on Inspection Rating when compared to Inspection Rating.

Component	Training Accuracy	Testing Accuracy	MSE	r^2	MSE Reduction
	Linear Regression				
D3010 Energy Supply	0.266	0.279	139.868	0.279	0.683
D3020 Heat Generating Systems	0.214	0.288	205.712	0.288	0.838
D3030 Cooling Generating Systems	0.185	0.144	286.895	0.144	0.866
D3040 Distribution Systems	0.241	0.266	200.328	0.266	0.909
D3050 Terminal & Package Units	0.232	0.220	235.094	0.220	0.867
D3060 Controls & Instrumentation	0.274	0.174	236.150	0.174	0.909
D3090 Other HVAC Systems	0.366	0.047	336.675	0.047	0.800
	Random Forest				
D3010 Energy Supply	0.599	0.266	142.478	0.266	0.677
D3020 Heat Generating Systems	0.541	0.324	195.341	0.324	0.846
D3030 Cooling Generating Systems	0.483	0.184	273.419	0.184	0.872
D3040 Distribution Systems	0.565	0.416	159.300	0.416	0.928
D3050 Terminal & Package Units	0.547	0.355	194.373	0.355	0.890
D3060 Controls & Instrumentation	0.688	0.420	165.794	0.420	0.936
D3090 Other HVAC Systems	0.701	0.512	172.547	0.512	0.898
	Gradient Boosting				
D3010 Energy Supply	0.471	0.367	122.784	0.367	0.722
D3020 Heat Generating Systems	0.352	0.381	178.847	0.381	0.859
D3030 Cooling Generating Systems	0.323	0.240	254.575	0.240	0.881
D3040 Distribution Systems	0.394	0.414	159.989	0.414	0.927
D3050 Terminal & Package Units	0.386	0.356	194.269	0.356	0.890
D3060 Controls & Instrumentation	0.571	0.441	159.752	0.441	0.938
D3090 Other HVAC Systems	0.707	0.427	202.325	0.427	0.880

The second set of models developed were also predictive regression models but aimed at predicting a Variability Rating based on several input parameters. Ultimately, Age-Based Obsolescence, Expected Rating, Section Age, Component, and Inspector were selected as input variables. The resulting models were measured for performance and fit by calculating values for training accuracy, testing accuracy, MSE and r-squared from the actual dataset, shown in **Table 10**. The results show that the linear regression is a very poor model but Random Forest and Gradient

Boosting regression techniques both resulted in very low MSE with high r-squared values indicating strong model fit. Of the two, Random Forest was better across all metrics.

Table 10. Results of predictive model on Variability Rating.

Regression Technique	Training Accuracy	Testing Accuracy	MSE	r^2
Linear	0.770	-1.37E16	8.46E16	-1.37E16
Random Forest	0.971	0.881	0.735	0.881
Gradient Boosting	0.817	0.821	1.101	0.821

Visual results of the Random Forest and Gradient Boosting Regression Models are shown in **Figure 17**. Looking at both plots, the Random Forest Regression appears to be more concentrated along the actual Variability Ratings with less outliers. The Gradient Boosting shows a more robust distribution with increased outliers. Both distributions show a slight overestimation at lower Variability Ratings (lower variability) and an underestimation at higher Variability Ratings (higher variability).

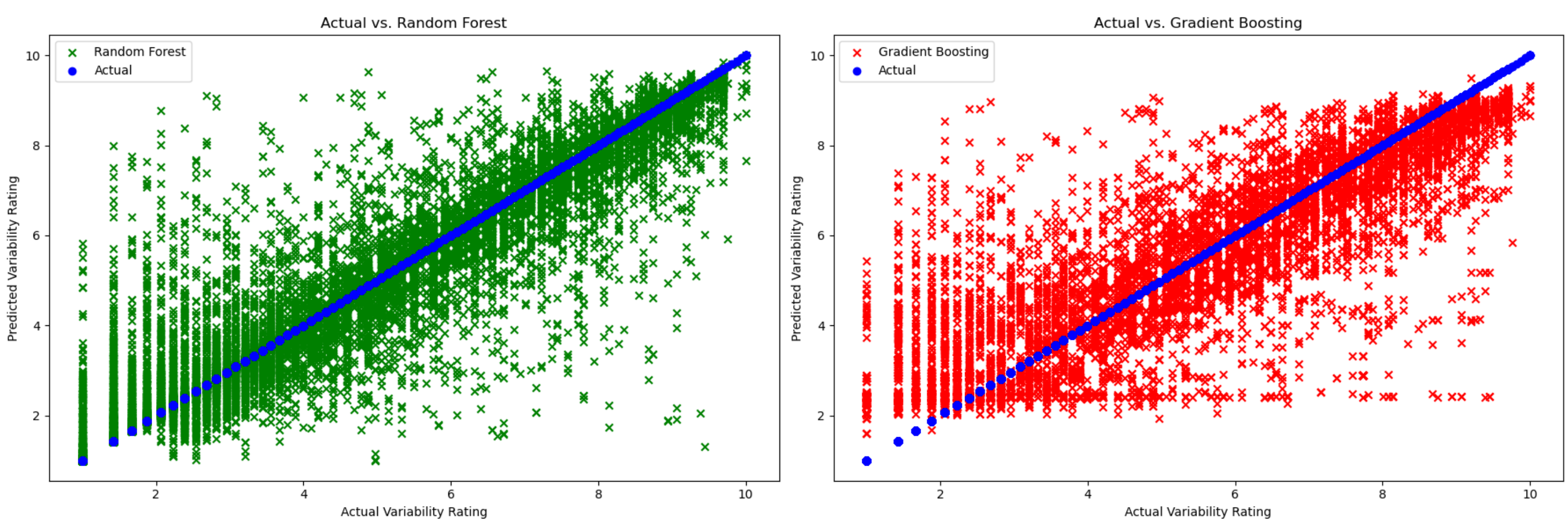

Figure 17. Predicted and Actual Variability Ratings for Random Forest and Gradient Boosting Regression.

To validate the results of the Variability Ratings predictive models k-fold cross validation was implemented. Using a 5-fold technique, the dataset was split into 5 components and tested individually and averaged to determine an overall accuracy for each regression technique. The results of this cross validation are shown in **Table 11**. Based on these results the Random Forest regression model was the best overall with an average training accuracy of 97.1%, average MSE of 0.735, and average r^2 of 0.881 showing overall good fit and performance.

Table 11. Results of Variability Rating cross validation.

Regression Technique	Parameter	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Linear	Training Accuracy	0.771	0.772	0.770	0.772	0.774	**0.772**
	Test MSE	1.408	1.49E+18	1.398	1.37E+18	1.491	**5.71E+17**
	Test R2	0.769	-2.43E+17	0.773	-2.23E+17	0.760	**-9.32E+16**
Random Forest	Training Accuracy	0.971	0.971	0.970	0.971	0.970	**0.971**
	Test MSE	0.736	0.737	0.713	0.724	0.763	**0.735**
	Test R2	0.880	0.879	0.884	0.883	0.878	**0.881**
Gradient Boosting	Training Accuracy	0.817	0.818	0.817	0.818	0.819	**0.818**
	Test MSE	1.127	1.140	1.122	1.122	1.160	**1.134**
	Test R2	0.815	0.814	0.818	0.817	0.813	**0.816**

Discussion and Conclusion

This study's findings highlight the prevailing challenges and research efforts within the field of HVAC inspections. Notably, manual inspections, a prominent practice in the domain, are susceptible to issues of subjectivity and cognitive bias. These challenges manifest as variability in findings due to differences in judgment and evaluation among inspectors. Additionally, inspection quality is influenced by factors such as experience, education on evaluation procedures, and inspection frequency. These are elements that literature has consistently recognized as pivotal for achieving reliable inspection results. Furthermore, the review underscores the growing role of technology and automation in HVAC inspections, a theme reflected in our research. There is potential for technological advancement but the limitations of autonomous robots and Unmanned Aerial Vehicles (UAVs) in intrusively inspecting HVAC systems in operation is still prevalent.

In the realm of improving HVAC condition determination, this study's analysis dovetails with strategies discussed in the literature. The deployment of probabilistic models, soft computing, and machine learning approaches is consistent with this paper's focus on predictive modeling. The importance of enhancing the accuracy of HVAC condition assessments is supported in literature, however, this study highlights the critical issue of variability and uncertainty in inspections that the existing literature has yet to fully address.

The key findings derived from a comprehensive analysis of an HVAC inspection dataset are presented, shedding light on the implications and potential future research directions. The results provide valuable insights into the complexities of HVAC inspection processes and the performance of predictive models. Data analysis revealed notable aspects of the HVAC inspection dataset. Visual observation showed that variables such as Section RDL and Section Age, exhibited normal distributions, while others, including RSL, Section CI, Age-Based Obsolescence, and Percentage Error, displayed significant skewness. This insight is crucial for understanding the inherent data structure and model development. Moreover, correlation analysis revealed strong relationships between Expected Rating and Age-Based Obsolescence, Percentage Error, Section Age, Section RDL, Squared Error, and Variability Rating using calculation of Spearman's Rho.

In the context of predictive modeling, analysis yielded noteworthy results. When predicting Inspection Ratings based on select input variables, including Age-Based Obsolescence, Section RDL, Section Age, and Component, a substantial reduction in Mean Squared Error was achieved.

However, low r^2 values suggested that these models had limited correlation with the actual data. When comparing the predicted Inspection Ratings to the Actual Expected Ratings, these models to be less reliable due to their low r-squared values and MSE reduction of 10 percent or less. This indicates that the overall dataset is not valid for predicting Inspection Ratings that can strongly aid in decision support when compared to Expected Ratings.

Conversely, in the second set of models aimed at predicting Variability Ratings, comprising Age-Based Obsolescence, Expected Rating, Section Age, Component, and Inspector as input variables, promising results were observed. Both Random Forest and Gradient Boosting regression techniques achieved notably low MSE and high r-squared values, indicating a strong model fit. Particularly, Random Forest outperformed Gradient Boosting with high accuracy, low MSE, and a near 0.881 r^2 when subject to cross validation. These models could play a pivotal role in evaluating the variability in inspection outcomes, a critical consideration in HVAC maintenance.

These findings suggest that predicting Inspection Ratings based on this dataset may not be feasible. The variations in model performance across regression techniques underscore the complexity of the task and the importance of future research in model enhancement and feature selection. On the other hand, the predictive models for Variability Rating show great promise in assessing the variability in inspection outcomes. Future research could focus on improving the accuracy of these models and their practical application in HVAC maintenance. Additionally, research into implementing predictive models as decision support tools for infrastructure inspection should also be pursued. This study lays the groundwork for addressing these challenges and improving the consistency and reliability of HVAC inspection processes.

Conclusion

The two interconnected research papers presented in this thesis collectively aim to enhance the efficacy of infrastructure asset management systems by addressing human inspection variability. The first paper, which broadly identifies variability in the context of BUILDER SMS within the Air Force, and the second paper, which focuses on HVAC systems, have jointly shed light on the challenges of human inspection and provided valuable insights into strategies for mitigating these challenges.

The first paper primarily addressed the goal of quantifying the variability in human inspection, using the BUILDER SMS infrastructure asset management program within the Department of Defense's infrastructure management as an example. In response to the research questions, it was found that significant variability exists within the human inspection processes, with the potential to impact the effectiveness of infrastructure asset management systems (Gordan et al. 2022). Furthermore, the study quantified the extent of variability in human inspection processes, using statistical analysis, demonstrating its potential to influence maintenance investment decisions. Notably, the findings indicated that MEP systems, specifically HVAC experienced highest inspection variability in many attributes when compared to other infrastructure components.

The second paper centered on HVAC systems, chosen as a critical testing bed due to their substantial impact on energy consumption and facility operations and based on findings from the first paper. Addressing the research questions, this paper determined that various strategies exist to reduce human inspection uncertainty in HVAC systems (Cai et al. 2023). The findings emphasized the importance of advanced technologies and automated inspections as promising avenues for improvement. Additionally, factors such as work experience, training, education, and demonstrating tasks, all played important roles in reducing inspection variability (Ball et al. 2017; Liu et al. 2023; Torabi et al. 2021). Importantly, this research culminated in the development of a model designed to predict a Variability Rating for HVAC system inspections, demonstrating the potential for enhancing reliability and reducing variability in HVAC inspections.

The overarching objective of these studies was to improve infrastructure asset management systems by addressing the issues associated with human inspection variability (Zhe et al. 2019). This variability in human inspections can significantly affect the allocation of resources and decision-making processes (Cai et al. 2023; Tharanga et al. 2022). The findings underscore the need for a more holistic approach that leverages advanced technologies and insights to reduce human inspection uncertainty. Additionally, the development of a model for establishing a Variability Rating in HVAC systems inspections provides a decision support tool for reducing variability.

In summary, these two papers collectively offer insights into the importance of addressing human inspection variability within infrastructure asset management systems (Gordan et al. 2022). The findings underscore the need for more standardized and technology-driven approaches to ensure infrastructure assets are efficiently maintained, thereby extending their lifespan, and optimizing resource allocation. As infrastructure asset management becomes increasingly critical, these studies offer a valuable step towards achieving more reliable and consistent inspection practices (Tharanga et al. 2022).

References

Agnisarman, S., S. Lopes, K. Chalil Madathil, K. Piratla, and A. Gramopadhye. 2019. "A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection." *Autom. Constr.*, 97: 52–76. https://doi.org/https://doi.org/10.1016/j.autcon.2018.10.019.

Akoglu, H. 2018. "User's guide to correlation coefficients." *Turkish J. Emerg. Med.*, 18 (3): 91–93. https://doi.org/https://doi.org/10.1016/j.tjem.2018.08.001.

Artusi, R., P. Verderio, and E. Marubini. 2002. "Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval." *Int. J. Biol. Markers*, 17 (2): 148–151. SAGE Publications Ltd STM. https://doi.org/10.1177/172460080201700213.

ASHRAE. 2018. *Standard Practice for Inspection and Maintenance of Commercial Building HVAC Systems*.

Aydin, C., and B. Ozerdem. 2006. "Air leakage measurement and analysis in duct systems." *Energy Build.*, 38 (3): 207–213. Elsevier.

Ball, G., E. Siemsen, and R. Shah. 2017. "Do plant inspections predict future quality? The role of investigator experience." *Manuf. Serv. Oper. Manag.*, 19 (4): 534–550. INFORMS.

Bartels, L. 2014. "Alternate methods to obtain facility condition assessment data using non-engineering resources." *Int. J. Inf. Electron. Eng.*, 4 (4): 326. IACSIT Press.

Besiktepe, D., M. E. Ozbek, and R. A. Atadero. 2021. "Condition Assessment Framework for Facility Management Based on Fuzzy Sets Theory." *Buildings*.

Bhanot, K., S. K. Peddoju, and T. Bhardwaj. 2018. "A model to find optimal percentage of training and testing data for efficient ECG analysis using neural network." *Int. J. Syst. Assur. Eng. Manag.*, 9 (1): 12–17. https://doi.org/10.1007/s13198-015-0398-7.

"BUILDER Sustainment Management System." 2012. *U.S. Army Eng. Res. Dev. Cent.* Accessed December 1, 2023. https://www.erdc.usace.army.mil/Media/Fact-Sheets/Fact-Sheet-Article-View/Article/476728/builder-sustainment-management-system/.

Cai, W., L. Huang, and Z. Zou. 2023. "Actively-exploring thermography-enabled autonomous robotic system for detecting and registering HVAC thermal leaks." *Autom. Constr.*, 152: 104901. https://doi.org/https://doi.org/10.1016/j.autcon.2023.104901.

Campbell, L. E., R. J. Connor, J. M. Whitehead, and G. A. Washer. 2021. "Human factors affecting visual inspection of fatigue cracking in steel bridges." *Struct. Infrastruct. Eng.*, 17 (11): 1447–1458. Taylor & Francis. https://doi.org/10.1080/15732479.2020.1813783.

Chan, A. P. C., D. W. M. Chan, and J. F. Y. Yeung. 2009. "Overview of the application of 'fuzzy techniques' in construction management research." *J. Constr. Eng. Manag.*, 135 (11): 1241–1252. American Society of Civil Engineers.

Dubey, P., J. Jackman, G. E. Kremer, and P. Kremer. 2023. "A Probabilistic Model to Estimate Automated and Manual Visual Inspection Errors BT - Flexible Automation and Intelligent Manufacturing: The Human-Data-Technology Nexus." K.-Y. Kim, L. Monplaisir, and J. Rickli, eds., 685–695. Cham: Springer International Publishing.

Fisk, W. J., W. Delp, R. Diamond, D. Dickerhoff, R. Levinson, M. Modera, M. Nematollahi, and

D. Wang. 2000. "Duct systems in large commercial buildings: physical characterization, air leakage, and heat conduction gains." *Energy Build.*, 32 (1): 109–119. Elsevier.

Gordan, M., O. Z. Chao, S.-R. Sabbagh-Yazdi, L. K. Wee, K. Ghaedi, and Z. Ismail. 2022. "From Cognitive Bias Toward Advanced Computational Intelligence for Smart Infrastructure Monitoring ." *Front. Psychol.* .

Grussing, M. N. 2014. "Life cycle asset management methodologies for buildings." *J. Infrastruct. Syst.*, 20 (1): 4013007. American Society of Civil Engineers.

Hegazy, T., S. Singh Ahluwalia, and M. Attalla. 2010. "Two condition indicators for building components based on reactive-maintenance data." *J. Facil. Manag.*, 8 (1): 64–74. Emerald Group Publishing Limited.

Herrera, G. J., C. A. Stokes, V. Peña, and S. V Howieson. 2017. *Review of the BUILDER Application for Assessing Federal Laboratory Facilities*. JSTOR.

Huang, Y., Y. Liu, C. Li, and C. Wang. 2019. "GBRTVis: online analysis of gradient boosting regression tree." *J. Vis.*, 22 (1): 125–140. https://doi.org/10.1007/s12650-018-0514-2.

Humplick, F. 1992. "Identifying error-generating factors in infrastructure condition evaluations." *Transp. Res. Rec.*, 106. NATIONAL ACADEMY OF SCIENCES.

Jing, G., X. Qin, H. Wang, and C. Deng. 2022. "Developments, challenges, and perspectives of railway inspection robots." *Autom. Constr.*, 138: 104242. https://doi.org/https://doi.org/10.1016/j.autcon.2022.104242.

Kadam, V., S. Kanhere, and S. Mahindrakar. 2020. "Regression Techniques in Machine Learning

& Applications: A Review." *Int. J. Res. Appl. Sci. Eng. Technol.*, 8 (10): 826–830. https://doi.org/https://doi.org/10.22214/ijraset.2020.32019.

Kim, I.-H., S. Yoon, J. H. Lee, S. Jung, S. Cho, and H.-J. Jung. 2022. "A Comparative Study of Bridge Inspection and Condition Assessment between Manpower and a UAS." *Drones*.

Kute, J. A., A. A. Pawar, and A. M. Kate. 2015. "Experimental analysis of performance of HPHX by varying the input parameters for HVAC application." e-ISSN.

Lamm, K. R., J. D. Delorit, M. N. Grussing, and S. J. Schuldt. 2022. "Improving Data-Driven Infrastructure Degradation Forecast Skill with Stepwise Asset Condition Prediction Models." *Buildings*.

Lan, L., P. Wargocki, and Z. Lian. 2011. "Quantitative measurement of productivity loss due to thermal discomfort." *Energy Build.*, 43 (5): 1057–1062. Elsevier.

Lei, X., Y. Chen, M. Bergés, and B. Akinci. 2021. "Formalized control logic fault definition with ontological reasoning for air handling units." *Autom. Constr.*, 129: 103781. Elsevier.

Liu, P., Y. Shi, R. Xiong, and P. Tang. 2023. "Quantifying the reliability of defects located by bridge inspectors through human observation behavioral analysis." *Dev. Built Environ.*, 14: 100167. https://doi.org/https://doi.org/10.1016/j.dibe.2023.100167.

Ma, Z., Y. Ren, X. Xiang, and Z. Turk. 2020. "Data-driven decision-making for equipment maintenance." *Autom. Constr.*, 112: 103103. https://doi.org/https://doi.org/10.1016/j.autcon.2020.103103.

Mathaudhu, S. S. 1997. "HVAC for energy showcase." *ASHRAE J.*, 39 (7): 27. American Society of Heating, Refrigeration and Air Conditioning Engineers, Inc.

Matos, R., H. Rodrigues, A. Costa, C. Requejo, and F. Rodrigues. 2023. "Facility manager perspectives on building condition assessment for Portuguese higher education institutions." *Build. Res. Inf.*, 51 (8): 949–964. Routledge. https://doi.org/10.1080/09613218.2023.2204415.

McDonald, S. J. 2023. "The Role of Facility Management Metrics in Promoting Sustainable Facility Management Practices Using BUILDER SMS: A Quantitative Causal Comparative / Correlational Study." *ProQuest Diss. Theses*. United States -- California: Northcentral University PP - United States -- California.

McKight, P. E., and J. Najab. 2010. "Kruskal-Wallis Test." *Corsini Encycl. Psychol.*, 1.

Moradi, S., T. Zayed, and F. Golkhoo. 2019. "Review on Computer Aided Sewer Pipeline Defect Detection and Condition Assessment." *Infrastructures*.

NADCA. 2021. *The NADCA standard for assessment, cleaning, and restoration of HVAC systems*.

Patro, S., and K. K. Sahu. 2015. "Normalization: A preprocessing stage." *arXiv Prepr. arXiv1503.06462*.

Pratt, C. 2023. "Human Inspection Variability in Infrastructure Asset Management: A Focus on HVAC Systems." Virginia Tech.

Rastegari, A., and M. Mobin. 2016. "Maintenance decision making, supported by computerized maintenance management system." *2016 Annu. Reliab. Maintainab. Symp.*, 1–8.

Ricketts, R. 2016. "Impact of large building airtightness requirements." *Proc. Therm. Perform. Exter. Envel. Whole Build. Int. Conf. Clear. FL, USA*, 4–8.

Schober, P., C. Boer, and L. A. Schwarte. 2018. "Correlation Coefficients: Appropriate Use and Interpretation." *Anesth. Analg.*, 126 (5).

See, J. E., C. G. Drury, A. Speed, A. Williams, and N. Khalandi. 2017. "The Role of Visual Inspection in the 21st Century." *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 61 (1): 262–266. SAGE Publications Inc. https://doi.org/10.1177/1541931213601548.

Seger, C. 2018. *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*.

Sein, S., J. Matos, and J. Idnurm. 2019. *Uncertainty in condition prediction of bridges based on assessment method – case study in Estonia*.

Sharma, K., A. Kumar, and K. Sharma. 2023. "Integration of Robotics and Artificial Intelligence in Non-Destructive Testing: A Review of Applications and Challenges." *e-Journal Nondestruct. Test.*, 28 (9). https://doi.org/https://doi.org/10.58286/28457.

Stallard (Voelker), M. M., C. A. MacKenzie, and F. E. Peters. 2018. "A probabilistic model to estimate visual inspection error for metalcastings given different training and judgment types, environmental and human factors, and percent of defects." *J. Manuf. Syst.*, 48: 97–106. https://doi.org/https://doi.org/10.1016/j.jmsy.2018.07.002.

Synovec, T., I. Howard, and L. Priddy. 2023. "Thirty-Year Airfield Pavement Behavioral and Economic Simulation to Improve Budget Allocation for the US Air Force Network." *J.*

Infrastruct. Syst., 29 (1): 5022008. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000727.

Tekin, H., I. C. Yilmaz, S. Atabay, and A. P. Gurgun. 2023. "IDENTIFICATION OF DEFECTIVE CONSTRUCTION WORKS DURING BUILDING INSPECTIONS." *Proc. Int. Struct. Eng. Constr.*, 10. https://doi.org/10.14455/ISEC.2023.10(1).QUA-01.

Tharanga, D. S. D., S. Sujeeva, and T. Huu. 2022. "Effect of Runtime on the Deterioration of HVAC Components in Building Services." *J. Infrastruct. Syst.*, 28 (1): 4021049. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000656.

Thomasson, R. O., and S. K. Sinha. 2015. "Asset Management: Performance, Sustainability, and Resiliency Model Development." *Pipelines 2015*, Proceedings.

Torabi, N., H. Burak Gunay, and W. O'Brien. 2021. "A review of common human errors in design, installation, and operation of multiple-zone VAV AHU systems." *J. Phys. Conf. Ser.*, 2042 (1): 12130. IOP Publishing. https://doi.org/10.1088/1742-6596/2042/1/012130.

USACE. 2017. *Army BUILDER SMS Inventory and Assessment Guide D30 HVAC*.

Uzarski, D. R., M. N. Grussing, and C. J. B. 2007. "Knowledge-Based Condition Survey Inspection Concepts." *J. Infrastruct. Syst.*, 13 (1): 72–79. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)1076-0342(2007)13:1(72).

Uzarski, D. R., M. N. Grussing, and B. Mehnert. 2018. *Knowledge-Based Condition Assessment Reference Manual for Building Component Sections*.

Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, basics, and computing of exploratory data* *analysis*. Duxbury Press.

Wang, Y., and J. Su. 2014. “Automated defect and contaminant inspection of HVAC duct.” *Autom. Constr.*, 41: 15–24. https://doi.org/https://doi.org/10.1016/j.autcon.2014.02.001.

Washer, G. A., M. M. Hammed, P. Jensen, and R. J. Connor. 2020. “Quality of Element-Level Bridge Inspection Data.” *Transp. Res. Rec.*, 2674 (2): 252–261. SAGE Publications Inc. https://doi.org/10.1177/0361198120907382.

West, J. S., S. W. Foster, and C. Jones. 2022. “Corrosion Condition Assessment Program for Maritime Asset Management.” *AMPP Annu. Conf. + Expo*.

Wilson, H., and D. Goldfein. 2019. *US Air Force Infrastructure Investment Strategy*.

Woodcock, K. 2014. “Model of safety inspection.” *Saf. Sci.*, 62: 145–156. Elsevier.

Yu, Y., D. Woradechjumroen, and D. Yu. 2014. “A review of fault detection and diagnosis methodologies on air-handling units.” *Energy Build.*, 82: 550–562. Elsevier.

Zhang, L., M. Leach, Y. Bae, B. Cui, S. Bhattacharya, S. Lee, P. Im, V. Adetola, D. Vrabie, and T. Kuruganti. 2021. “Sensor impact evaluation and verification for fault detection and diagnostics in building energy systems: A review.” *Adv. Appl. Energy*, 3: 100055. Elsevier.

Zhe, S., T. Pingbo, S. Ying, and X. Wen. 2019. “Visual-Semantic Alignments for Automated Interpretation of 3D Imagery Data of High-Pier Bridges.” *Comput. Civ. Eng. 2019*, Proceedings.

Appendix A

Final 9 – Facility System Quick View Report

Basic Description:

A cross-tab report (pivot table); for each Facility this report displays the System CI for each of the 22 Systems.

If this report is tabular, then what are the rows?

Each row is a unique Building/Facility.

Is this report filtered? If so, how?

Yes, this report only shows Facilities where at least one System has been entered into BUILDER and it only includes facilities where the BUILDING STATUS is not equal to either "Demolished" or "Transferred".

Additional Notes:

"Installation Code" is the Site Number; "Installation Name" is the Site Name; "Special Area" is the Complex Name.

Report Columns:

ORDER	COLUMN NAME
1	*Installation Code*
2	*Installation Name*
3	*Special Area*
4	*RPUID*
5	*Bldg Num*
6	*Bldg Name*
7	*MDI*
8	*Category Code*
9	*English UM*
10	*Metric UM*
11	*Size (English)*
12	*Floors*
13	*Const Year*
14	*BCI*
15	*A10*
16	*A20*
17	*B10*
18	*B20*
19	*B30*
20	*C10*
21	*C20*
22	*C30*
23	*D10*
24	*D20*
25	*D30*
26	*D40*
27	*D50*
28	*E10*
29	*E20*
30	*F10*
31	*F20*
32	*G10*
33	*G20*
34	*G30*
35	*G40*
36	*G90*

QA 12C – Inspection Matrix – By Site, Inspector, System

Basic Description:

A cross-tab report (pivot table); for each Inspector, a count of Green, Amber, Red scores by System.

If this report is tabular, then what are the rows?

Each row is a unique INSPECTOR broken down by System.

Is this report filtered? If so, how?

No.

Report Columns:

ORDER	COLUMN NAME
1	Site No
2	Inspector
3	Site Name
4	System
5	Total # of Inspections
6	# of Green Scores
7	# of Amber Scores
8	# of Red Scores
9	% of Green Scores
10	% of Amber Scores
11	% of Red Scores

Additional Notes:

Analysis of Inspector scores can help reveal any inspector biases or inconsistencies.

QC 6 – Inspections Report

Basic Description:

A list of all inspections entered into BUILDER.

If this report is tabular, then what are the rows?

Each row is a unique inspection.

Is this report filtered? If so, how?

No, this report shows ALL inspections, not only the LATEST.

Report Columns:

ORDER	COLUMN NAME
1	Site No
2	Site Name
3	Complex No
4	Complex Name
5	Facility No
6	Facility Name
7	Const Year
8	Bldg Size
9	Floors
10	System
11	Component
12	Mat / Equip Type
13	Section Subtype
14	Section Name
15	Section Qty
16	Section UM

Report Columns, continued:

ORDER	COLUMN NAME
17	Section Year
18	Section Year Source
19	Section DL
20	Section Age
21	Section RDL
22	Section RSL
23	Painted?
24	Paint Year
25	CSCI
26	CSCCI
27	Section Comments
28	Insp Date
29	Insp Type
30	Insp Source
31	Inspector
32	Expected Rating
33	Comp Rating
34	Paint Rating
35	Num Inspection Images
36	Insp Comments
37	Site ID
38	Complex ID
39	Facility ID
40	System ID
41	Component ID
42	Section ID
43	Inspection ID