Construction of a Historical Infrastructure Price Index

Ibraheem Saqib Ellahi, Christopher Lynch, Abhik Saha, Eden Sobalvarro, Jesse Valdes

Dr. Carl Kitchens

Abstract

This research project aims to develop a comprehensive historical construction price index spanning from the 20th century onwards, recognizing significant shifts influenced by factors such as inflation, technological advancements, and efficiency improvements.

In the early 20th century, the pricing of American infrastructure construction lacked digitization. To address this, microfilms from the Engineering News-Record are digitized through scanning microfilm and processing to make them machine readable. Leveraging microfilm provides access to a historical journal with weekly editions dating back to the 19th century, enabling an examination of prices for various construction elements, job-related salaries, and awarded contracts. After processing the aggregate microfilm data, images are corrected for transcription errors, and weights are assigned to individual projects. The organized aggregate data is then categorized at the city-year-infrastructure type level.

The extraction process employs text parsing and image formatting techniques to unveil relevant construction pricing information. This involves identifying monthly awarded construction contracts based on regional parameters. Specifically, machine learning methods, including Amazon Textract and Python data scraping syntax, are utilized to efficiently extract construction pricing from thousands of pages at a time.

The findings from this project hold the potential to assist policymakers and those involved in constructing new buildings in estimating potential costs. By identifying trends among historical decisions, this information contributes to more informed decision-making regarding future construction expenses.

Background Information

In the initial half of the 20th century, information regarding the pricing of American infrastructure construction is scarce. This is because data from the first half of the 20th century is not digitized and are stored in microfilm from the Engineering News-Record.

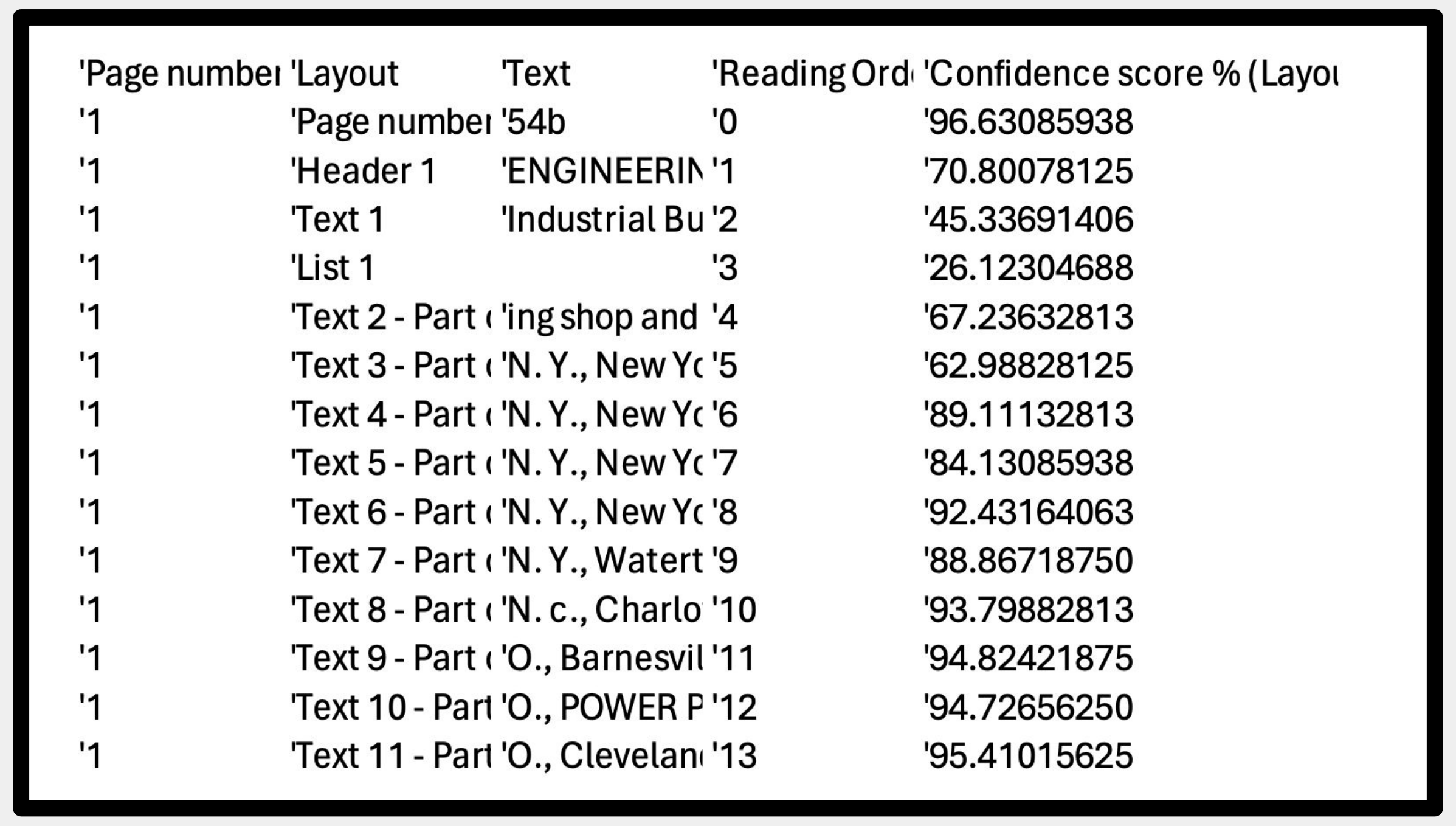
To address this scarcity of data needed to create a historical construction price index, the microfilm has to be digitized and analyzed to assist in developing a historical construction price infrastructure index, which is used to measure the changes in cost of infrastructure over.

The index can give insights on the changes of infrastructure construction prices throughout the 21st century. By connecting the first half of the 20th century of data to the second half of the 20th century, the index provides insights into the changes in infrastructure construction prices across the United States.

Data Collection

Specific construction pricing scanned from microfilm machine pictured above.

Amazon Texttract is used to extract data and numerical values from the page into a Excel File pictured below.



'Page number	'Layout	'Text	'Reading Ord	'Confidence score % (Layou
'1	'Page number	'54b	'0	'96.63085938
'1	'Header 1	'ENGINEERIN	'1	'70.80078125
'1	'Text 1	'Industrial Bu	'2	'45.33691406
'1	'List 1		'3	'26.12304688
'1	'Text 2 - Part (	'ing shop and	'4	'67.23632813
'1	'Text 3 - Part (	'N. Y., New Yo	'5	'62.98828125
'1	'Text 4 - Part (	'N. Y., New Yo	'6	'89.11132813
'1	'Text 5 - Part (	'N. Y., New Yo	'7	'84.13085938
'1	'Text 6 - Part (	'N. Y., New Yo	'8	'92.43164063
'1	'Text 7 - Part (	'N. Y., Watert	'9	'88.86718750
'1	'Text 8 - Part (	'N. c., Charlo	'10	'93.79882813
'1	'Text 9 - Part (	'O., Barnesvil	'11	'94.82421875
'1	'Text 10 - Part	'O., POWER P	'12	'94.72656250
'1	'Text 11 - Part	'O., Clevelan	'13	'95.41015625

Methods

In order to create a database, two major stages are involved, directly tying into each other: data collection and data processing.

1. Data Collection: Data was collected through microfilm in Strozier and Online PDFs on HathiTrust.org from the "Engineering News Record."
   - i. Specific construction pricing information was scanned from the articles.
   - ii. The clarity of old microfilm was optimized to enhance the accuracy of data processing.
2. Data Processing: The processed data was scanned through Optical Character Recognition (OCR).
   - i. Amazon Textract was utilized due to its broad formatting options, essential for handling the diverse nature of the data.
   - ii. The technique was employed to sort the mass amounts of non-digitized data into a usable form.

Preliminary Findings

Through this project, data revealing the evolution of construction prices over the years has been uncovered. The historical journal under examination consistently presents figures pertaining to material and labor costs of significant construction projects each year. Machine learning tools, such as Amazon Textract and Python, enable the transformation of data from graphs and charts into readable text, facilitating a clearer understanding of numerical changes.

As the project is ongoing, its future relies on utilizing data analysis methods to determine the actual rate of change between the years through the microfilm data extraction process. Ongoing efforts will involve continuous data exploration to enhance the project, offering a more comprehensive representation of changes over time. The future findings aim to enable more effective and efficient construction projects, providing economists with the means to draw insightful comparisons.

References

Boughton , V. T., et al., editors. "Engineering News-Record 1936." *Engineering News-Record*, vol. 117, 1936.

"OCR Software, Data Extraction Tool - Amazon Textract - AWS." *Amazon Extract*, Amazon, aws.amazon.com/textract/. Accessed 21 Feb. 2024.