

# A Bus Delay Prediction Model Based on Machine Learning Algorithms

Carl Ka To Ma, Hugo Fang, and Cato Yuan Wen under supervision of Mr. Yiming Wu  
Shanghai High School International Division, Shanghai, China

**Abstract**—Bus delay prediction is a practical feature made available in many smart cities, allowing people to estimate their travel time when taking public transportation. Existing algorithms are built on the prerequisite that every vehicle in the bus network is equipped with GPS capable of sharing real-time location. This is not the case in developing cities. Therefore, the aim of this paper is to introduce a bus delay prediction model that only requires common, accessible data like date and time, weather, bus route, and traffic incident reports. While these broad categories of data have little meaning, they are preprocessed into specific factors that show a much stronger correlation with bus delay, such as peak or off-peak hours, visibility, precipitation, etc. A total of 24 input features are passed into a feed-forward neural network, trained with data collected in Toronto, Canada, over the past six years. While Toronto is used as an example because it provides convenient access to open data, the model is designed to work in all cities around the world. The performance of this algorithm is satisfactory: its 81% accuracy, despite being far from the near-100% accuracy archived by algorithms utilizing real-time GPS location, is fairly acceptable in a developing city before advanced systems are installed.

**Index Terms**—Bus delay, delay prediction, machine learning, neural network, open data, smart cities

## I. INTRODUCTION

The emergence of smart cities fueled by advances in Artificial Intelligence (AI) changed the way people live over the past decade. From optical character recognition to automated driving systems, AI has brought countless benefits, simplifying human tasks or even achieving ones otherwise impossible by humans. One prerequisite of AI, however, is large sets of training data. This problem is hopefully solved in smart cities, where data is collected from the infinitely many sensors belonging to the Internet of Things and utilized to improve the quality of city services. Furthermore, according to the European Commission, a smart city is one with an efficient transport network, an upgraded water supply, a more interactive and responsive city administration, and so on. The focus of this essay will be on the first characteristic—transportation.

### A. Background

Transportation has been revolutionized in smart cities with the introduction of applications like Google Maps, which provide a variety of services including route planning, real-time traffic information, and more. These services are made possible by the joint efforts of the mapping service provider (e.g. Google Maps) and the local transport department, responsible for the technology and the access to data, respectively. In some developed cities, the two offices are able to

collaborate successfully, enhancing their services and bringing added convenience to the public. One example of a successful implementation is in the city of Hong Kong (HKSAR). Not only does Google Maps return a feasible route when the user enters a starting point and destination, it also considers factors like the bus schedule and potential delays if one chooses to take public transportation. According to Fabrikant (2019) from Google Research, bus delay predictions adopted by Google Maps rely on a sophisticated machine learning model of which details are not disclosed. As a result, an additional literature review is conducted to examine existing pieces of research and methodologies on the implementation of bus delay predictions.

### B. Literature Review

The significance of this topic and the potential benefits it brings urge researchers worldwide to conduct similar explorations. In terms of algorithms, early researchers like Angelo (1999) and Patnaik (2004) used statistical models—namely, nonlinear time series and multivariable regression. Each of these early algorithms has their own limitations when applied to the prediction of bus delays, as Yin, Zhong, et al. (2006) have summarized: low accuracy, poor performance, or invalid assumptions. Yin and Zhong then suggested that the rapid development of machine learning enables the modeling of bus delays with support vector machines (SVM) and artificial neural networks (ANN). Indeed, these machine learning algorithms are becoming the spotlight of discussion. Sjaafjell, Dahl, and Skogen (2014) proposed means to optimize the SVM and ANN models, eventually reaching a conclusion that their implementation of the SVM model would outperform the ANN model with less training, while the opposite holds true given more training. In terms of input parameters, Patnaik (2004) collected data specifically for his multivariable regression model, including the exact number of passengers boarding or exiting the bus. Recent models featuring machine learning are capable of processing large datasets, so researchers are inspired to use the real-time location of buses as an input parameter, persuading cities to install GPS systems on every vehicle. Sjaafjell's SVM and ANN models, for example, analyze the data collected by the GPS trackers during the three previous trips.

### C. Objectives

Utilizing real-time GPS location as an input parameter in a bus delay prediction model, as Google Maps must have done

in developed cities like Hong Kong, would no doubt increase the model accuracy. However, an issue surfaces in developing cities, where buses are not equipped with GPS systems. Is it still possible to predict bus delays with machine learning in these cities with data other than the real-time location of buses? In this paper, a model that only takes into account the data at hand, data collected by existing nodes in the Internet of Things, will be discussed and evaluated. The objective of this project is therefore to create a generalized bus delay prediction model based on machine learning that can provide fairly reliable bus delay predictions, not just for developed cities, but, more importantly, for developing cities.

## II. EASE OF USE

### A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

## III. PREPARE YOUR PAPER BEFORE STYLING

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads— $\LaTeX$  will do that for you.

### A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

### B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m<sup>2</sup>” or “webers per square meter”, not “webers/m<sup>2</sup>”.

Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

- Use a zero before decimal points: “0.25”, not “.25”. Use “cm<sup>3</sup>”, not “cc”.)

### C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \quad (1)$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

### D. $\LaTeX$ -Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., (1)). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in  $\LaTeX$  will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

$\BIBTeX$  does not work by magic. It doesn’t get the bibliographic data from thin air but from .bib files. If you use  $\BIBTeX$  to produce a bibliography you must send the .bib files.

$\LaTeX$  can’t read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

$\LaTeX$  does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it’s supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won’t be any anyway) and it might stop a wanted equation number in the surrounding equation.

### E. Some Common Mistakes

- The word “data” is plural, not singular.

- The subscript for the permeability of vacuum  $\mu_0$ , and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
- Do not use the word “essentially” to mean “approximately” or “effectively”.
- In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
- Do not confuse “imply” and “infer”.
- The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the “et” in the Latin abbreviation “et al.”.
- The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

#### F. Authors and Affiliations

**The class file is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

#### G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style

provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

#### H. Figures and Tables

a) *Positioning Figures and Tables:* Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

TABLE I  
TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	Subhead	Subhead
copy	More table copy <sup>a</sup>		

<sup>a</sup>Sample of a Table footnote.



Fig. 1. Example of a figure caption.

**Figure Labels:** Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

#### ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

## REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

## REFERENCES

- [1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
- [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [4] K. Elissa, “Title of paper if known,” unpublished.
- [5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.*, in press.
- [6] Y. Yoroazu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [7] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.