

Gary Gregg
Poster Review Comments
DATA 591, Winter 2018
3 March 2018

Estimating Soil Moisture Using Multi-Spectral Satellite Imagery

I noticed that the word “using” is not capitalized, while the rest of the title words are. My training has been that only stop words, like articles (“a,” “an,” “the”) or prepositions (“of,” for example) should not be capitalized in titles. “Using” looks like a verb gerund, or present participle here, and should probably be capitalized. You probably will take care of this, but the “*Version 4: Live mockup*” in the title bar should be removed in your final version. Near the bottom of the first column, there is the word “*shapefiles*.” Was this intended as two words, “shape files?” Can the term be defined? In the second column, the bar plots of the mean absolute error per model type is a powerful graphic; I like it. In the third column, under “*Hydrological Model and Interpolation*,” there is the word “*kriging*.” Is that a word? Perhaps it could be put in italics, or defined, so that the reader knows it is not a misspelling. **This is a great project, and your results speak for themselves.**

Seattle Gentrification Atlas

In the section “*Project Objectives*” in the first column, there are the words “...*status of block groups neighborhoods*...” Is “groups” an adjective here, modifying “neighborhoods,” or is it a noun on its own? I think the “s” should be dropped from “groups.” In some cases, there is inconsistent use of periods to terminate bulleted, or numbered items. Most of these end in a period, but in some cases they do not. For example, numbered item (2) under “Conclusions” is missing a terminating period. I recommend consistency in this regard. **I really like your comparative maps showing how a different definition of gentrification results in a different color-coded map. This is powerful. Your poster had the best attention to detail than all that I looked at.**

Name Entity Disambiguation

In the first column, under “Method,” you note that the model is raw data in csv format “...as Zicong presented.” I think you probably mean that Zicong presented this at our initial results meeting at the end of January. However, our participants at the formal presentation will know nothing about this. I recommend dropping the words “as Zicong presented.” Also in this paragraph are the words “four step.” I think you meant “**fourth step.**” “r-step” as it begins a sentence should be capitalized. Same paragraph: “Then we will **so** some post processing.” You probably mean “Then we will **do** some post processing.” Same paragraph: Change the last sentence to, “...we will focus our work before **the** r-step.” **Do you really want to use future tense here?** Is the work done yet, or not? I recommend present tense, or past tense instead, assuming the work is done. Recommendation is either uniformly past tense, or uniformly present tense. Don’t mix them.

Your iterative cycle diagram with blue bubbles and arrows is good, and easily comprehensible. I like it.

Second column, “**Levenshtein**” appears to be misspelled. Change “*misspell names*” to “*misspelled names.*” Regarding the words, “...one string is identical to another, the only difference is in the order of the name units.” Break this into two sentences, on the comma.

I could not see the connection of the word cloud graphic in the third column to the **obvious value of the research.** Is it just notional? Could you clear that up?

Automatic Tagging and Generation of Marketing Messages

Your poster has more written text than any I looked at. That is neither good nor bad, just a matter of style. However, that may put an extra emphasis on correct punctuation, verb tense, and grammar. You might expect interested participants to spend a lot of time reading through your poster. Throughout the text, I noted inconsistent separation of sentences. This applies to both whitespace separation, and commas where periods should be. In the text that follows, I simply rewrote some of the sentences. I recommend finding the match for my substitution in the poster, and making the same, or a similar change.

First paragraph, under "Background": *"Amplero provides a Artificial Intelligence Marketing Platform..."* should be *"Amplero provides **an** Artificial Intelligence Marketing Platform..."* Also in the first paragraph, separate *"marketing technology stack and data sources"* with commas: *"marketing, technology, stack, and data sources."*

Second paragraph under *Background* – rewrite it thusly: *"Developing a system to automatically tag text messages enables Amplero to process large amounts of their client's future marketing messages."* Check throughout your document to be certain that each sentence begins with a capital letter. I found some exceptions. Change the last sentence under *Data* to: *"50% of the data in the tagging columns are missing."*

Under *Methodology*, change to: *"We then performed an overall statistics analysis on the data, including dimension analysis, missing values, and cluster distribution on each tag."* Throughout the document, check for sufficient whitespace between sentences. There should uniformly be one or two space characters between sentences. In *Part 1*, under *Methodology*, it appears that the paragraph ends in a question, but there is no question mark. In the second paragraph of the same section, it appears the sentence *"We only kept verbs, nouns and punctuation mark."* I think you meant *"punctuation **marks**."* Also, this sentence is separated from the one before it with a comma instead of a period. Rewrite the last two sentences in this paragraphs thusly: *"We applied filtering to words that appear frequently, but have little or no meaning. These include stop-words, unnecessary symbols, numbers and punctuation."*

Under “3. Modeling.” Change to: “The next step is to classify our text messages into predefined tags based on their content. We use Naive Bayes, Decision Tree, and Linear Support Vector Machines (SVM) to classify the preprocessed text messages.”

“We calculated the accuracy score...” → “We calculated the accuracy score, which is the portion of correctly classified messages to the total number of messages. We then compared the accuracy of the different classification models to make a selection.” Check to be sure your verb tense is either consistently present tense, or consistently past tense.

“To make the training data more structured...” → “To make the training data more structured, we removed the messy parts of each message, and replaced numbers with the corresponding variable.” I am not sure I even understood that sentence correctly, so you may wish to reword it. When you say “messy” part of each message, what do you mean?

“Here is an example of a message after cleaning...”

“We used an n-grams method to generate messages...” “...we used bigrams to generate a list of next words for each word.” **Your barcharts are all powerful results, and very eye-catching.**

Change: “Negative tone messages tend to be harder to predict while positive tone message are the easiest to categorize, and predict.”

Change: “Linear SVM has the best prediction performance on a ‘lead’ tag than Naive Bayes and Decision Tree.”

Change: “Different cleaning techniques were applied to the tagging part and the message generation parts of each message.”

Change: “We manually tagged the Tone and Lead part of the messages to fill the missing tagging columns.”

Customer Segmentation for Variable Annuity Products

I noticed you defined “Variable Annuity” (VA) more than once in the introduction. That may not be necessary. I think once you mention it once, you can use VA throughout the document with the expectation that the reader knows what you mean. Third sentence in the introduction, change to: *“As sales of variable annuities in the U.S. market grows, insurance companies face big challenges in terms of pricing their products due to the uncertainties of policyholders’ behaviors.”* The third sentence in the introduction is a monster run-on. Change to something like: *“This project offers predictive analysis of customers’ behaviors for our project sponsor, Milliman. Milliman is one of the world’s largest providers of actuarial and related products and services. We here provide the sponsor reasonable customer segmentation suggestions to differentiate profitability of policyholders for their client insurers. Our goal will be to guide insurers product design and marketing.”* Four sentences where there was one run-on before.

I like how the objectives are broken down into well-defined deliverables.

However, the sentences in the deliverables are crammed with difficult data science terminology. Will the readers nod in understanding, or will their eyes glaze over? Is it possible to rewrite these deliverables while explaining the terms, even if it means more words?

I would remove passive voice throughout. Example: *“Two logistic model with LASSO regularization were constructed...”* Change to: *“We constructed two logistic models with LASSO regularization...”*

Under *Cons* in the *Clustering Algorithm Comparison*: *“...might lead to mis grouping.”* Should this read *“...might lead to miscategorization...”*? If so, change “mis grouping” to “mischaracterization” throughout.

Under *Optimal Number of Clusters*: The text inside the blue arrows often does not fit within the arrow, and is therefore cut off.

The “*Conclusion*” section should probably be “*Conclusions*.” Change *“...difficult to scale to large data set...”* to *“difficult to scale to a large data set.”* Again, I see some passive voice in the document. Change *“KMeans was introduced”* to *“we introduced KMeans...”*

Your poster had more “meat” than any of the others in it in terms of graphics, and visualization of methods. Very good!

Amplero Time Series Prediction

This poster has better spelling, sentence structure and grammar than many of the others, however, the beginning sections are a little light in content. Perhaps you can break down the Background section with subsections for a problem statement, research questions, and objectives. **I like very much your pipeline path visualization, and the tabular results of your “Assessment” section.** I assume the Assessment section substitutes for the Conclusions section, which seems fine to me.