International Journal of Scientific Research and Management (IJSRM)
||Volume||07||Issue||02||Pages||AA-2019-01-09||2019||
Website: www.ijsrm.in ISSN (e): 2321-3418
Index Copernicus value (2016): 93.67, (2017):89.90, DOI: 10.18535/ijsrm/v7i2.aa01

Astronomy of the day is more effective than seasonal decomposition in modeling and predicting the rates of Amazon review misspellings

Renay Oshop,

www.AyurAstro.com

Abstract
This study shows that the misspelling rates in Amazon reviews can be effectively modeled as having cyclical components by using seasonal decomposition, and this effectiveness is almost identical to that gained by using a neural network that uses basic astronomical placements of the day. While the modeling of past Amazon misspelling rates was slightly more effective using a neural net and astronomy of day as compared to seasonal decomposition, future values were only effectively predicted by the neural net that used astronomy of day.

Introduction

Astrology tends to use multiple factors altogether such as Sun placement in a zodiac, as well as Moon placement and placements of the planets, as well as many other significations.

These multiple factors relate to each other as multiple dependent variables in math parlance. Such multivariate analysis is particularly amenable to modeling via neural networks. (Odom & Ramesh, 1993)

Despite the appropriateness and fit of neural networks to the astrological approach, a literature search shows that publication has been sparse on success or failure in prediction of real-world phenomena via astronomical features within such a neural network. (Kulkarni & S., 2012) (Karimbaev, 2017) This paper contributes to such literature.

The average misspelling rate for *Amazon.com* reviews on any given day may sound rather abstruse and without pattern. This paper asks: is there any cyclicity to average daily misspelling rates over time? If there is, does the cyclicity correspond to cycles in planetary placements?

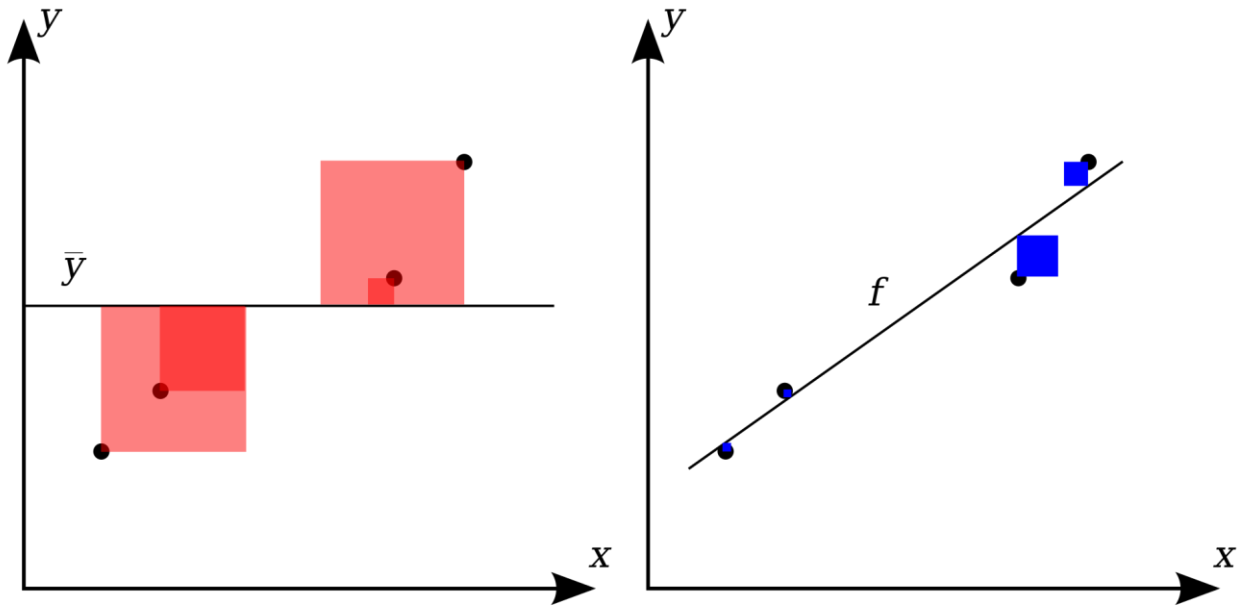
Seasonal decomposition is a standard method from commerce and science for detecting and modeling cyclicity (also known as seasonality) in time series. (Brownlee, 2017)

For this paper, while the modeling of past Amazon misspelling rates was slightly more effective just using astronomy of day and a neural network as compared to seasonal decomposition (R-squared of 0.74 as compared to an R-squared of 0.72), future values were only effectively predicted by using astronomy of day. (R-squared of 0.84 as compared to an R-squared of -0.32.)

The results make heavy use of R-squared, or the coefficient of determination, as a classic measure of fitness of different models. It is defined to be:

R^2 = 1 - (sum of squares of differences between a model's predicted values and observed values)/ (sum of squares of differences between the average value and observed values).

Let's see this pictorially. In the following image, four black dots represent four actual values to be modeled.



On the left, an average value (the horizontal black middle line) is the basis for the differences (a. k. a. residuals) that are depicted as squared via the red squares.

On the right, a linear regression model gives predictions as shown by the slanted black line. The differences (residuals) at the four points between the predicted values and the actual values are depicted as squared via the blue squares.

The smaller that the sum of areas of blue is relative to the sum of areas in red, the higher is the R-squared value.

The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R-squared is to 1. (Orzetto, 2010)

Thus, an R-squared of one is a sign of perfect prediction, because the areas of the blue squares have gone to zero, since the residuals of the model have gone to zero, and one minus zero is one. The model would receive an A+ grade, if you will. No higher R-squared can be obtained.

On the other hand, an R-squared of zero is only as good as guessing the average value. Since the red and blue squares are the same, their sums have a ratio of one, and one minus that ratio of one gives zero. The model could be said to have a C grade.

A model can also have a D grade or an F by having a negative R-squared value. This just means that the sum of the blue squares is even bigger than the sum of the red squares and is a mark of a poor model indeed.

Materials and Methods

Upon request, Stanford University's SNAP big data repository provided all 79,743,786 *Amazon.com* reviews up to mid-2014. (McAuley, n.d.) *Mathematica* software's astronomy resources provided the astronomical data. (Wolfram Research, Inc., 2019) That software was also used for the big data processing and making the graphs. (Oshop, 2019)

All other information about the reviews was discarded other than dates and texts of the review bodies.

A first, necessary, and difficult step was figuring out what counted as a misspelling in the Amazon reviews.

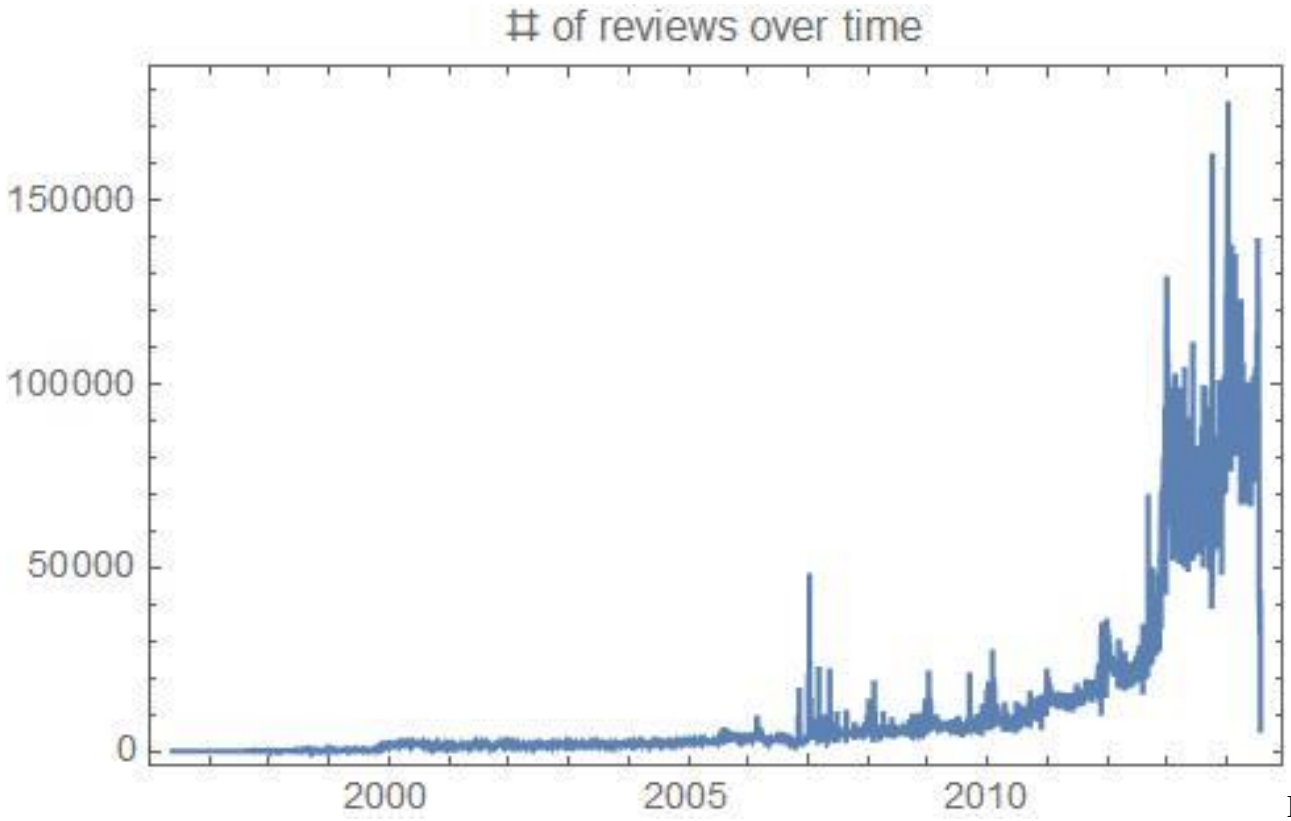
Just counting words that are not in a computerized dictionary would not be a good idea. After all, the reviews might include brand names, author names, international variants, URLs, emoji, slang, and other creative word play. Moreover, the sheer number of all those reviews made scanning them for this task computationally expensive.

Andrew Foss, an astrologer who holds a PhD in computer science, offered a very good solution: find the 100 most common, explicitly wrong words that are not in the computerized dictionary in a randomized 5% of the whole body of Amazon text. The following is a word cloud of these top 100 misspelled words in the random 5% of the Amazon corpus. The bigger the word is, the more often it was misspelled there.

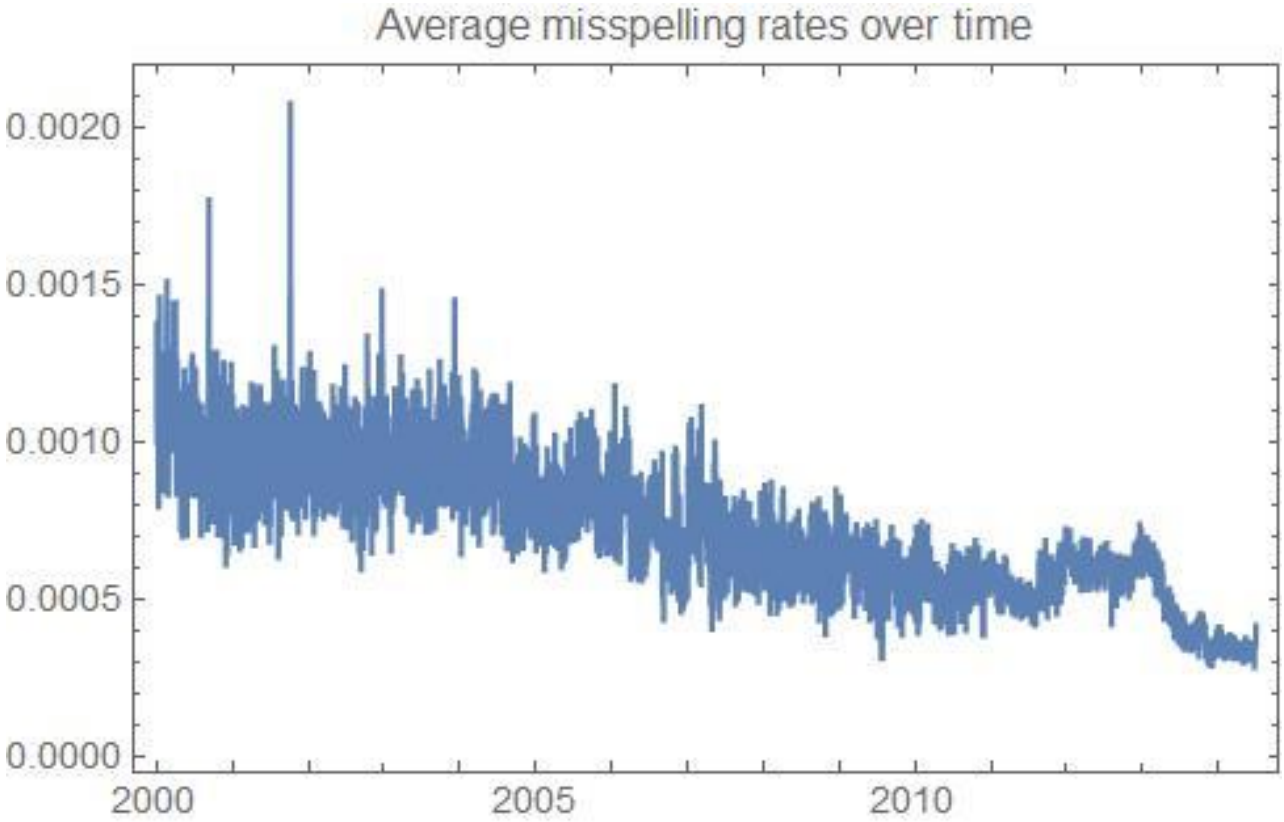
Each of the 79,743,786 reviews was then scanned again, checking first to ensure the review was in English and nonempty. Only 0.223% (or 177,513) of reviews were either empty or not in English. These were not included in the subsequent graphs or calculations.

For each review, counting just the marker words and dividing by the total number of words (defined by the blank spaces with emoji removed) gave a misspelling rate for that review. These rates were then averaged for all reviews in a day to give an average daily misspelling rate.

The following is a picture of the total number of the reviews over time. Note the drop off to zero in the number of reviews before 2000 and after mid-2014. Thus, only dates from January 1, 2000 to July 1, 2014 were considered further.



Here is a picture of the average daily rates of misspelling for the 100 wrong words over time.



Because of the diminutive nature of these values, understandable given that they are the daily misspelling rates of only 100 marker words, a transformation multiple of 10,000 was applied across all the values to make the images in the Results section below more readable. To retrieve the original values, one just needs to divide individual data points by 10,000. (The R-squared shown in the results are the same in either case.)

Various approaches to seasonal decomposition of the first 80% of misspelling rate values were taken. (Machine Learning Made Beautifully Simple for Everyone, n.d.) They are depicted below. AIC, AICc, and BIC are measures of information loss in the various models, with the lowest information loss given by the lowest of these values, but they are ranked by internal R-squared. All will be compared to the astronomy-based model.

Name	Description	AIC	AICc	BIC	R squared
M,N,N	Simple exponential smoothing with multiplicative error. Multiplicative error, no trend, and no seasonality model.	35360.6949	35360.7006	35379.7498	0.7223
M,Md,N	Damped trend exponential method. Multiplicative error, multiplicative damped trend, and no seasonality model.	35367.1064	35367.1263	35405.2161	0.7223
A,N,N	Simple exponential smoothing with additive errors. Additive error, no trend, and no seasonality model.	35810.9256	35810.9313	35829.9805	0.7223
A,A,N	Holt's linear method with additive errors. Additive error, additive trend, and no seasonality model.	35814.9171	35814.9312	35846.6751	0.7223
A,Ad,N	Damped trend linear method with additive errors. Additive error, additive damped trend, and no seasonality model.	35817.4651	35817.485	35855.5748	0.7223
M,A,N	Holt's linear method with multiplicative errors. Multiplicative error, additive trend, and no seasonality model.	35365.0839	35365.0981	35396.842	0.7221
M,Ad,N	Damped trend exponential method with multiplicative errors. Multiplicative error, additive damped trend, and no seasonality model.	35372.7891	35372.809	35410.8988	0.7215
M,A,M	Holt-Winters method with multiplicative seasonality. Multiplicative error, additive trend, and multiplicative seasonality model.	35455.2547	35456.1572	35728.374	0.7213
M,N,A	Multiplicative error, no trend, and additive seasonality model.	35466.4367	35467.2577	35726.8527	0.7211
A,A,A	Holt-Winters additive method with additive errors. Additive error, additive trend, and additive seasonality model.	35912.8345	35913.7369	36185.9537	0.7207
M,Ad,M	Holt-Winters damped method with multiplicative seasonality. Multiplicative error, additive damped trend, and multiplicative seasonality model.	35468.0353	35468.98	35747.5062	0.7206
A,N,A	Additive error, no trend, and additive seasonality model.	35910.3479	35911.1689	36170.7639	0.7206
M,A,A	Holt-Winters additive method with multiplicative errors. Multiplicative error, additive trend, and additive seasonality model.	35479.0612	35479.9636	35752.1805	0.7205
A,Ad,A	Holt-Winters damped method with additive seasonality. Additive error, additive damped trend, and additive seasonality model.	35917.8224	35918.767	36197.2932	0.7205
M,N,M	Multiplicative error, no trend, and multiplicative seasonality model.	35455.4831	35456.304	35715.8991	0.7203
M,M,M	Multiplicative error, multiplicative trend, and multiplicative seasonality model.	35461.3809	35462.2834	35734.5002	0.7203
M,Ad,A	Holt-Winters damped method with additive seasonality. Multiplicative error, additive damped trend, and additive seasonality model.	35490.7396	35491.6843	35770.2105	0.7201
M,Md,M	Multiplicative error, multiplicative damped trend, and multiplicative seasonality model.	35468.6897	35469.6344	35748.1606	0.7198
M,M,N	Exponential trend method. Multiplicative error, multiplicative trend, and no seasonality model.	35451.2817	35451.2959	35483.0397	0.6937

To compare the utility of using astronomical data, the astronomy of each day was calculated using astronomy, not astrology, tools. The complete set of factors included is: Sun, Moon, Mercury, Venus, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto equatorial right ascension placements, and those planets' apparent retrogression statuses. Lunar nodes were not included.

The astronomical right ascension (i.e., the astrological Tropical degree) of the planet, moon, or star at 00:00am, the start of that day UTC time, was used for all the days from January 1, 2000 to July 1, 2014.

(Sidereal placements with *Lahiri ayanamsha*, a popular alternative zodiac for which precession is accounted, were also modeled with an R-squared that was poorer by 2.4%.) Apparent retrogressions of the planets were also calculated for the start of day.

Midnight UTC at the start of day was chosen because the original Amazon review data gave the date but only 0:00 UTC as the time of each Amazon review with no further explanation. Surely, the reviews were written at other times too, but no finer granularity exists in the decades-long public data.

The astronomy data for a day and the misspelling rate for the day were joined together for all 5,296 days into a spreadsheet.

For the first 80% of data and astronomy values, an optimized neural network search was made which performed well with an internal R-squared of 0.73992 (+/- 0.00 standard deviation). (Machine Learning Made Beautifully Simple for Everyone, n.d.) Its astronomy fields and their importance are shown in the following table.

Field	**Importance**
Saturn degree of right ascension (0 to 360)	0.74214
Pluto degree of right ascension (0 to 360)	0.12837
Jupiter degree of right ascension (0 to 360)	0.07335
Neptune degree of right ascension (0 to 360)	0.03105
Saturn geocentric apparent retrogression status (prograde or retrograde)	0.00505
Uranus degree of right ascension (0 to 360)	0.00428
Mars degree of right ascension (0 to 360)	0.00388
Venus degree of right ascension (0 to 360)	0.00313
Pluto geocentric apparent retrogression status (prograde or retrograde)	0.00167
Mercury degree of right ascension (0 to 360)	0.00147
Moon degree of right ascension (0 to 360)	0.0013
Mercury geocentric apparent retrogression status (prograde or retrograde)	0.00129
Neptune geocentric apparent retrogression status (prograde or retrograde)	0.00107
Jupiter geocentric apparent retrogression status (prograde or retrograde)	0.00102
Sun degree of right ascension (0 to 360)	0.00081
Mars geocentric apparent retrogression status (prograde or retrograde)	0.00008
Uranus geocentric apparent retrogression status (prograde or retrograde)	0.00004

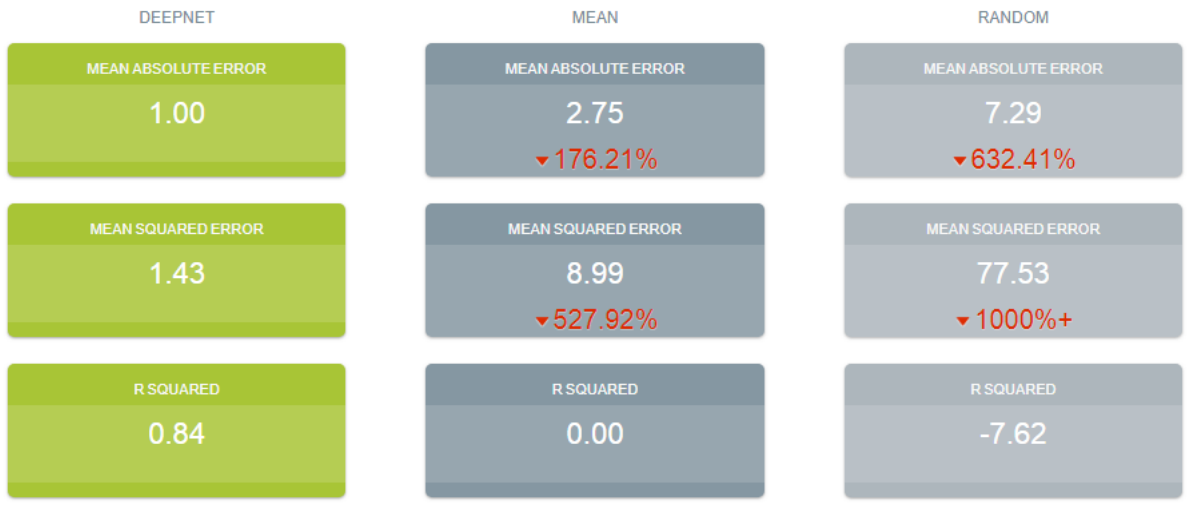
Now that we have the best methods for modeling the first 80% of the time series of Amazon review misspellings using methods that are unaccompanied and accompanied by astronomical data, we can compare the two approaches for efficacy in predicting the final 20% of future data.

Results

Here are the time series seasonal decompositions' results for predicting the last 20% of data. As you can see, the best performance still has only an R-squared of -0.032. (MAE, MSE, SMAPE, MASE, and MDA are various ways of characterizing the residuals.)

Name	MAE	MSE	R squared	SMAPE	MASE	MDA
M,M,M	1.08	1.48	-0.032	0.2189	2.11	0.4854
M,M,N	1.08	1.51	-0.0563	0.2179	2.1054	0.5335
M,N,M	1.13	1.52	-0.0603	0.23	2.2086	0.492
A,N,N	1.13	1.52	-0.0627	0.2304	2.2123	0.0009
M,N,N	1.13	1.52	-0.0634	0.2304	2.2129	0.0009
M,Md,N	1.13	1.52	-0.064	0.2305	2.2134	0.0604
A,Ad,N	1.13	1.52	-0.0644	0.2305	2.2137	0.0746
M,Ad,N	1.13	1.52	-0.0666	0.2307	2.2158	0.0642
A,N,A	1.13	1.53	-0.067	0.2306	2.2142	0.4873
M,Md,M	1.13	1.53	-0.067	0.2306	2.2148	0.4929
M,Ad,M	1.13	1.53	-0.0681	0.2307	2.2157	0.4816
M,N,A	1.13	1.53	-0.0687	0.2307	2.2153	0.4816
A,Ad,A	1.13	1.53	-0.0722	0.231	2.2185	0.4873
M,Ad,A	1.14	1.54	-0.0754	0.2313	2.2211	0.4797
M,A,N	1.21	1.74	-0.2169	0.2445	2.3598	0.4674
A,A,N	1.07	1.81	-0.268	0.2141	2.103	0.5335
M,A,M	1.09	1.85	-0.2923	0.2167	2.1247	0.491
M,A,A	1.12	1.93	-0.351	0.2248	2.1875	0.4873
A,A,A	1.19	2.09	-0.4651	0.2425	2.3203	0.4816

Compare those R-squared to the following for the predictive ability of the optimized neural net that is based on astronomy values.



This chart displays, in the usual artificial intelligence industry way, very good results. The residuals or errors in predictions for the test 20% group by the neural net (a. k. a. Deepnet) in green are dramatically smaller than the other standard methods of prediction in gray, based on either the mean (average) rate of the test 20% data or an approach using random data points therein. Moreover, the strong R-squared of 0.84 suggests good correlation of predicted misspelling rates to actual future values only for the astronomical data of the neural net. Note that the adjusted R-squared is also 0.84.

Discussion

Future misspelling rates in Amazon reviews were successfully predicted using only basic astronomy data in a neural network. The standard tool of modeling a time series through seasonal decomposition fared very poorly in comparison.

Ultimately, the daily misspelling rates in Amazon reviews are a human expression, making the very high R-squared given by the astronomy-based model all the more remarkable, as the social sciences tend to show much lower R-squared than the more physical sciences. (A concise guide to market research: The process, data, and methods using IBM SPSS Statistics, 2014)

Astronomy of a day can give real, predictive ability in a set of human behavior.

These results encourage and substantiate the hope that the new tool of machine learning can lead us further to other objective realms of human experience. Its use with astrology may in fact offer a bridge to understanding subjective human states, perhaps making them subjective no more.

References

[1] (2014). In E. Mooi, & M. Sarstedt, *A concise guide to market research: The process, data, and methods using IBM SPSS Statistics* (p. 211). New York: Springer. doi:10.1007/978-3-642-12541-6

[2] Brownlee, J. (2017, January 30). *How to Decompose Time Series Data into Trend and Seasonality*. Retrieved January 14, 2019, from MachineLearningMastery.com: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

[3] Karimbaev, T. (2017, Dec 2). *Astrological Predictions Checked by Machine Learning.* Retrieved from DZone: https://dzone.com/articles/astrological-prediction-of-bitcoin-price-checked-b

[4] Kulkarni, P., & S., S. S. (2012). Use of Neural Netwoks in Horoscope Prediction. *International Conference on Recent Trends in Engineering & Technology.* Chandwad.

[5] *Machine Learning Made Beautifully Simple for Everyone*. (n.d.). Retrieved January 14, 2019, from BigML: https://bigml.com

[6] McAuley, J. (n.d.). *Amazon Product Data*. Retrieved January 14, 2019, from http://jmcauley.ucsd.edu/data/amazon/

[7] Odom, M., & Ramesh, S. (1993). A Neural Network for Bankruptcy Prediction. In *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real-World Performance.* Probus Publishing Company.

[8] Orzetto. (2010, September 6). *Coefficient of determination*. Retrieved January 20, 2019, from Wikimedia: https://commons.wikimedia.org/wiki/File:Coefficient_of_Determination.svg

[9] Oshop, R. (2019, January 11). *Studying Amazon Review Misspellings with Respect to Astronomy*. Retrieved January 14, 2019, from https://wolfr.am/APah8n9h

[10] Wolfram Research, Inc. (2019). *Mathematica* (Version 11.2 ed.). Champaign, IL, USA: Wolfram Research, Inc.