Analysis of NSIDC Dataset Downloads and Metadata

The Relationship Between Extant Metadata and Download Rates of Datasets in the National Snow and Ice Data Center Repository

Yulia Kolesnikova
University of Washington
Information School
Seattle, WA
kolesy@uw.edu

Adam Lathrop
University of Washington
Information School
Seattle, WA
adl5@uw.edu

Bree Norlander
University of Washington
Information School
Seattle, WA
norlab@uw.edu

An Yan
University of Washington
Information School
Seattle, WA
yanan15@uw.edu

Abstract

Few research studies have quantitatively analyzed metadata elements associated with scientific data reuse. By using metadata and dataset download rates from the National Snow and Ice Data Center, we address whether there are key indicators in data repository metadata that show a statistically significant correlation with the download count of a dataset and whether we can predict data reuse using machine learning techniques. We used the download rate by unique IP addresses for individual datasets as our dependent variable and as a proxy for data reuse. Our analysis shows that the following metadata elements in NSIDC datasets are positively correlated with download rates: year of citation, number of data formats, number of contributors, number of platforms, number of spatial coverage areas, number of locations, and number of keywords. Our results are applicable to researchers and professionals working with data and add to the small body of work addressing metadata best practices for increasing discovery of data.

Keywords

metadata, data curation, machine learning, data citation, data reuse, data repositories

1. Introduction

The complexity and volume of science data has been increasing at an unprecedented rate, enabling numerous opportunities for data to be re-analyzed by others (Hanson, Sugden, & Alberts, 2011; Overpeck, Meehl, Bony, & Easterling, 2011). Hey, Tansley, and Tolle contend that we are in a new era of research defined by Jim Gray as the “fourth paradigm”: “almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, ‘data-intensive’ science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other" (p. xxx, 2009). Science has become more collaborative and open, with more data sharing than ever before. The reuse of data generates new scientific insights, stimulates innovations across fields, and allows for the verification of original results (Tenopir et al., 2011).

Researchers such as Tenopir et al. (2011; 2015), Kim et al. (2013), and Uhlir (2010) have documented the benefits of data reuse, however there is a dearth of research into what factors influence the reuse of scientific data. Zimmerman (2007) conducted an interview-based study and found that data quality and availability of certain types of data contribute to greater data reuse in the ecology research community. Piwowar and Vision (2013) conducted one of the few quantitative assessments of data reuse. They compiled a dataset with 9724 instances of in-text citations of GEO or ArrayExpress accession numbers which they used as their independent variable. The results of their multivariate regression analysis showed a nine-percent increase in citations for publications that deposited data in an open data repository. Belter (2014) used citation counts of three datasets from the National Oceanographic Data Center to study the impact of data curation and found a relationship between citation rates, the year a dataset was versioned, and the discipline that cited the dataset. However, he did not perform in-depth quantitative analysis into metadata factors that impact data citation counts or data reuse.

We were specifically interested in how metadata affects data reuse. Metadata has long been considered essential for the discovery of resources. Qin, Ball, and Greenberg (2012) state that "metadata is the foundation for data discovery, use, and preservation" (p. 62) and "metadata for scientific data can be considered as mission-critical in scientific data discovery, use, and citation" (p. 64). However, to our knowledge, there are no published studies that employ statistical methods to identify the metadata factors that influence data reuse, or predict future data reuse.

By using metadata and dataset download rates (as a proxy for dataset reuse, see section 2.1) for several hundred datasets from the National Snow and Ice Data Center (NSIDC) data repository, we quantitatively addressed the following:

1) What are key indicators in data repository metadata that show a statistically significant correlation with the download count of a dataset?
2) Can we predict data reuse via machine learning?

Our results are applicable to the work of data producers, data curators, and data management professionals particularly in the tasks of determining metadata best practices and data ingest procedures.

2. Data

The NSIDC defines the scope of datasets within their repository as "cryospheric data and data from programs or instruments deemed of importance to the cryospheric community" (NSIDC, n.d.). We used a subset of these datasets and their associated metadata as input to supervised machine learning models. Our dependent variable was the count of downloads per dataset from unique IP addresses, which we used as a proxy for potential dataset reuse. The initial phase of the research was to gather metadata for each dataset and explore the data. The second phase of the research involved identifying potential factors associated with higher download rates (potential reuse) using machine learning algorithms.

2.1 Defining Reuse

We chose download rates as our measure of reuse, however, (re)usage metrics can vary wildly in definition, be it by dataset views, dataset download rates, or citation rates in peer-reviewed research. We used a convenience selection of dataset download rates provided by the NSIDC for two main reasons: 1) the data was made available to us and 2) we had a limited time period in which to conduct our study. Future studies could employ other metrics for data reuse.

It is important to recognize that download rate is not equivalent to reuse rate. However, download metrics are often easy to access and we use them as a proxy for indicating the *potential* for reuse. Download rates do not suffer from the time-lag associated with publication, and take into account data use that may not lead to publication. Future studies could verify the correlation between download rates and published reuse.

Several NSIDC datasets included in our study are dynamic in that they are updated regularly. We recognize that these may in fact be downloaded regularly by the same researchers whose internet infrastructure includes regular IP address changes. However, we found that by removing datasets that are updated daily or yearly, our linear regression results did not significantly change, so we felt it safe to leave these datasets in for data analysis (see section 3.3 for further discussion).

2.2 Data Gathering and Pre-Processing

Upon our request, on April 13, 2016, NSIDC User Services kindly provided us with a dataset consisting of 820 dataset IDs (unique identifiers) and the number of downloads from unique IP addresses associated with each of the dataset IDs. The download rates served as the dependent variable in our supervised learning methods. The dataset IDs for each dataset were then used to scrape associated metadata provided on the NSIDC's public website. The combination of scraped metadata and the download rates were then used for the core of this analysis.

We scraped the following metadata elements from the NSIDC's publicly available website on May 18th, 2016:

- Data Format

- Contributors[1]
- Spatial Coverage
- Spatial Resolution
- Temporal Coverage
- Temporal Resolution
- Parameters, Platforms
- Sensors
- Dataset Version
- Dataset DOI
- Citation Date (Year)
- Locations
- Keywords
- Creation Date
- Last Update Date
- Dataset Title

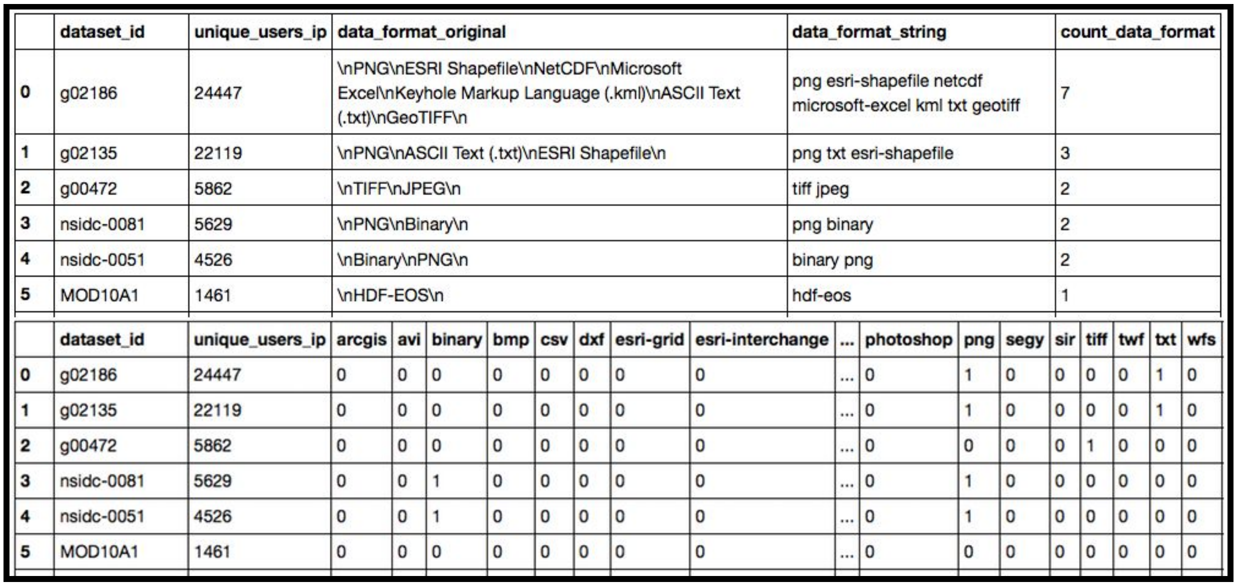
On May 22, 2016 we scraped additional data including:

- Links to NSIDC and external websites
- Links to other NISDC datasets
- Dataset IDs referenced in the “See Also” section
- Links to the given dataset
- Dataset ID numbers for datasets that link to the given observation

To scrape the elements, we wrote a Python (v. 2.7.11, Python Software Foundation, 2015) script which utilized the modules BeautifulSoup (v. 4.4.1, Richardson, 2015), Requests (v. 2.10.0, Reitz, Benfield, & Cordasco, 2016), Numpy (v. 1.10.2, van der Walt, Colbert, & Varoquaux, 2011), Pandas (v. 0.17.1, McKinney, 2010), and Re (RegularExpressions) (v. 2.2.1, Python Software Foundation, 2015). We then combined the scraped data into a Pandas dataframe with the associated NSIDC dataset IDs and download rates and saved in .csv file format for further cleaning, analysis, and to enable us to share the same raw dataset.

While we started with 820 datasets, we subsequently removed all rows in which all scraped metadata elements were empty, reducing our dataset to 797 rows. We cleaned the data and then separated the following elements into unique dataframes of “dummy” variables (binary variables that indicate the presence or absence of a given term): contributors, data formats, keywords, locations, sensors, platforms, and update frequency. We also counted the number of elements and added new features of counts (see Figure 1).

[1] We did scrape contributor names and performed analysis using these names. However, we felt that without the permission of the contributors, we could not include names in our analysis paper or in a publicly available dataset. The contributor names are publicly available on the NSIDC website, but will not be available with this paper or any associated datasets available for public consumption.



	dataset_id	unique_users_ip	data_format_original	data_format_string	count_data_format
0	g02186	24447	\nPNG\nESRI Shapefile\nNetCDF\nMicrosoft Excel\nKeyhole Markup Language (.kml)\nASCII Text (.txt)\nGeoTIFF\n	png esri-shapefile netcdf microsoft-excel kml txt geotiff	7
1	g02135	22119	\nPNG\nASCII Text (.txt)\nESRI Shapefile\n	png txt esri-shapefile	3
2	g00472	5862	\nTIFF\nJPEG\n	tiff jpeg	2
3	nsidc-0081	5629	\nPNG\nBinary\n	png binary	2
4	nsidc-0051	4526	\nBinary\nPNG\n	binary png	2
5	MOD10A1	1461	\nHDF-EOS\n	hdf-eos	1

	dataset_id	unique_users_ip	arcgis	avi	binary	bmp	csv	dxf	esri-grid	esri-interchange	...	photoshop	png	segy	sir	tiff	twf	txt	wfs
0	g02186	24447	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	1	0
1	g02135	22119	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	1	0
2	g00472	5862	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0
3	nsidc-0081	5629	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0
4	nsidc-0051	4526	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0
5	MOD10A1	1461	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Figure 1: Top half of image represents the original scraped data (data_format_original) from the metadata element "Data Format," followed by the cleaned data in string format (data_format_string), and finally the count of data formats per dataset (count_data_format). The bottom half of the image represents the sparse dummy-variable dataframe created from data_format_string.

2.3 Summary of Data

Prior to performing data analysis, we explored in detail the individual variables. The full list of variables and variable types are listed in the appendix in Table A1. The specific data types were chosen to facilitate modeling and analysis via our planned techniques. Certain variables in our dataset, for instance "scrape_date," and "doi_address_clean," were not used directly in data analysis, but retained to understand how the data was generated, for record-keeping purposes, and to assess the data for accuracy when merging dataframes.

2.4 Exploratory Analysis

We performed exploratory data analysis to look for potential patterns in the data, understand outliers, identify data for cleaning, formulate hypotheses, and determine whether further data collection was necessary. Summary statistics of the dependent variable and the ordinal variables in our dataset are shown in Table 1.

Feature (per observation)	Min	Max	Mean	Median	SD
Download Rate (Dependent Variable)	1	24447	148.25	37	1196.87
Version	1	34	2.48	1	6.18
Count of Data Formats	0	8	1.33	1	0.88
Count of Sensor Types	1	27	1.78	1	1.8

Count of Platform Types	1	17	1.73	1	1.58
Count of Contributors	1	10	2.6	2	1.65
Count of Spatial Coverage Areas	0	5	1.13	1	0.43
Count of Spatial Resolutions	0	3	0.39	0	0.59
Count of Named Locations	0	10	2.34	2	1.56
Count of Keywords	0	64	10.12	8	8.11
Count of Reference Links to NSIDC & External Webpages	0	7	1.63	1	1.48
Count of Reference Links to Other NSIDC Datasets	0	8	0.25	0	0.89
Count of Other Datasets that Reference the Given Dataset	0	7	0.27	0	0.87

Table 1: Central tendency for original features

2.5 Model Building

The distribution of download rates per dataset is strongly right-skewed (see Figure 2). While a handful of datasets experienced thousands of downloads, the vast majority of datasets were downloaded fewer than 100 times.

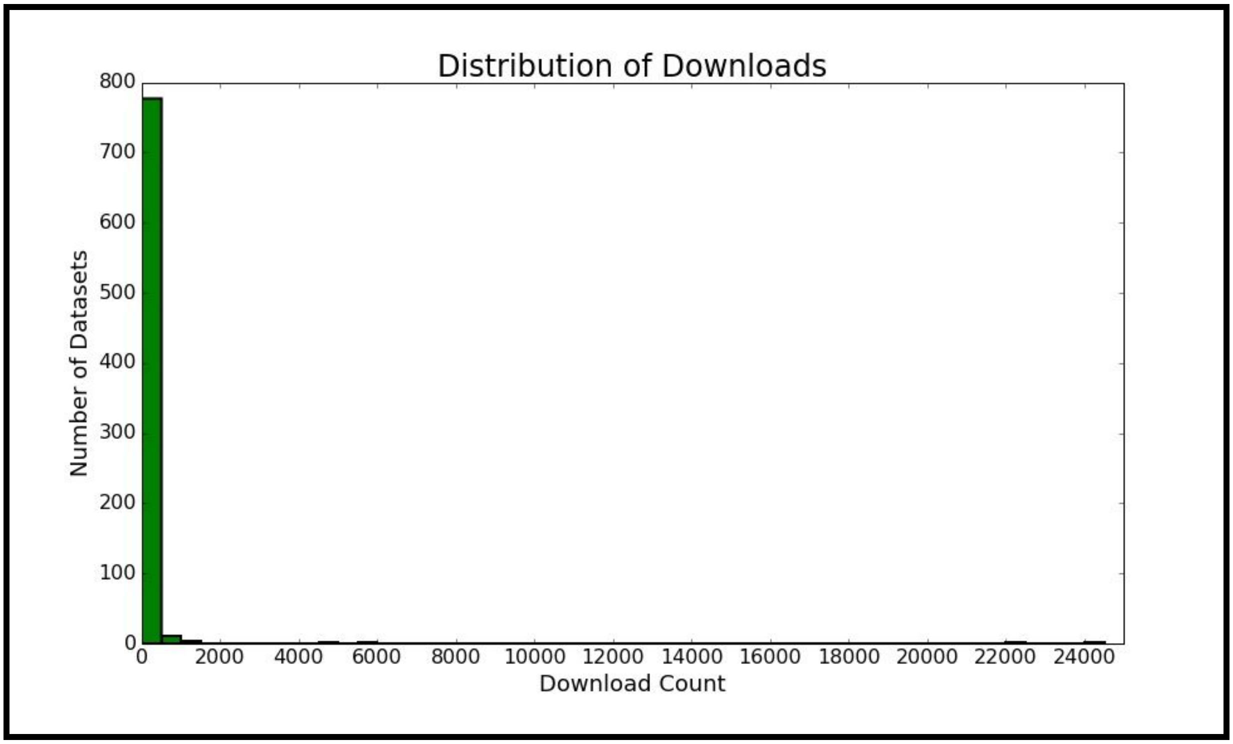


Figure 2: Distribution of download counts for the full dataset.

We employed a logarithmic scale (log natural), as recommended for social science research by Gelman and Hill (2007), for the download rate to better approximate a normal distribution. Figure 3 clearly shows that the vast majority of datasets fall within the midrange of download rates. Because the $\log_e$ distribution is closer to a normal distribution, we can approximate a more linear trend in the data (see Figure 4).

Before modeling, we split the full dataset into training and testing sets using an 80/20 split. Each author used their own training and testing sets but set random seeds for reproducibility. All models were trained on the training sets and predicted outcomes were tested on the testing sets. We describe the modeling process for each technique in detail below.

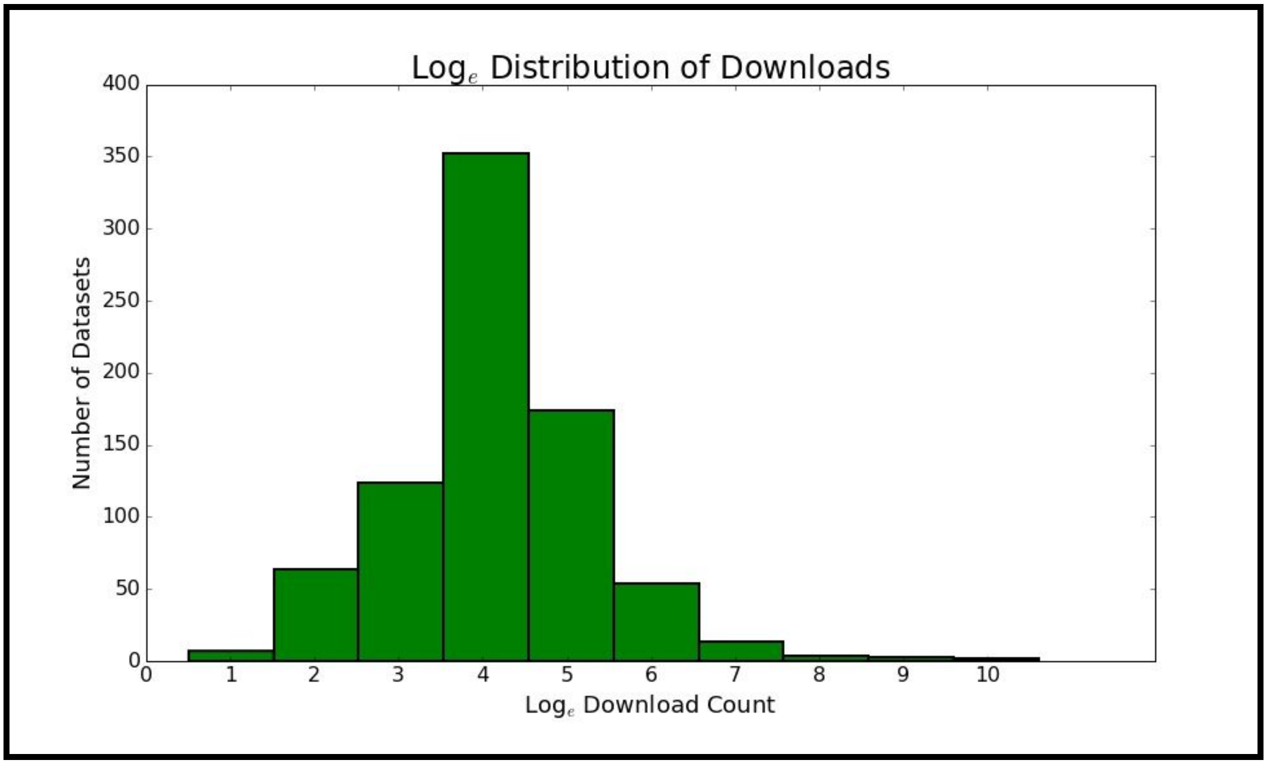


Figure 3: Log-normal distribution of dataset downloads

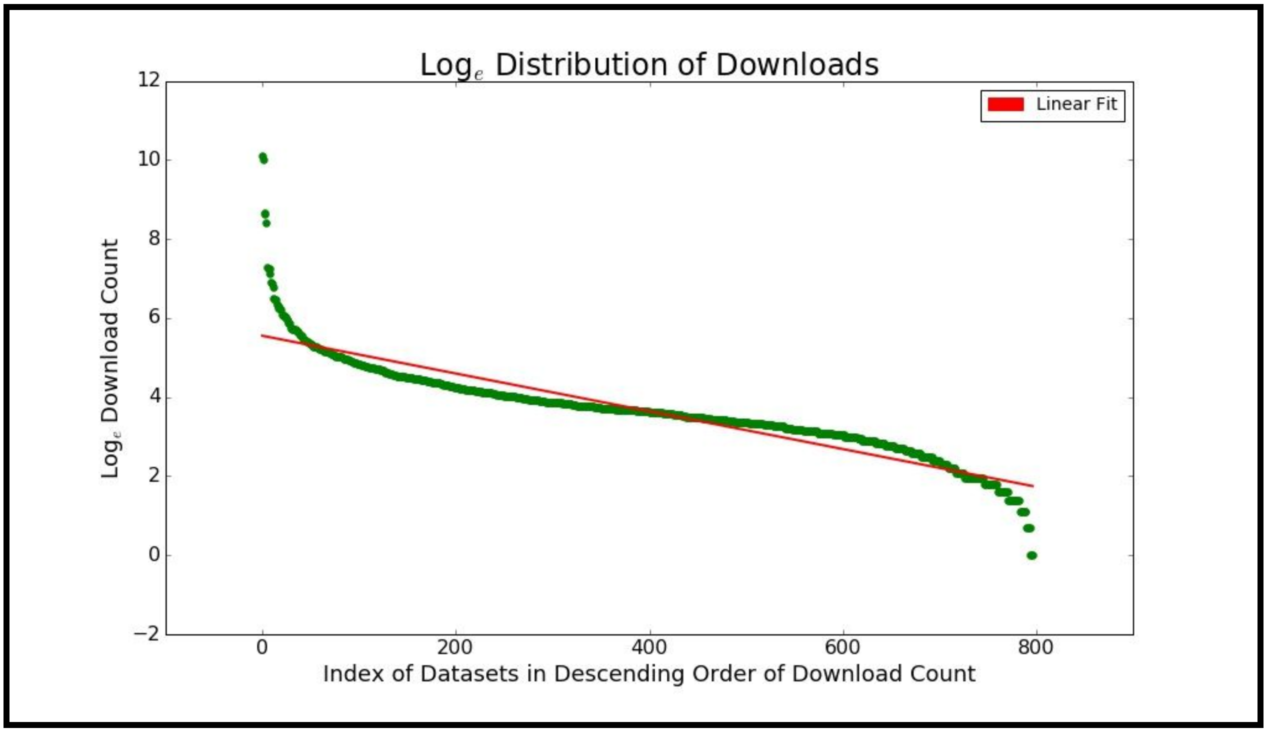


Figure 4: Log-normal distribution of dataset as a better fit for linear modeling.

Decision Trees

There are several advantages to using decision trees for prediction and for determining feature importance, such as their ease of modeling and interpretation, their ability to extract important features from data, and their ability to make use of data that may not fit a normal distribution. Because our dependent variable was not binary, we chose to use regression trees rather than classification trees.

We created one decision tree model using R (version 3.2.4) and the RPart package version 4.1.10 (Milborrow, 2014). We initially used the following independent variables (before pruning): data format count, citation year, spatial coverage count, sensor count, update frequency, and keyword count. We pruned the tree using RPart's cross-validated error rate in the Complexity Parameter Table.

Additionally, we created several regression trees using Python (version 2.7.11) and the Scikit-learn module version 0.17 (Pedregosa et al., 2011) with a maximum number of leaf nodes of 10 (chosen based on ease of reading and RMSE values). In these models we regressed the $\log_e$ of the download counts on our dummy variables (see Section 2.2).

Linear Regression

We also employed linear regression modeling since the $\log_e$ of download count follows a fairly linear trend (Figure 4), and linear regression is both simple and easily interpretable. Linear regression also indicates which independent variables may be highly correlated with the dependent variable based on p-value results. We used the standard p-value threshold of significance ($p < 0.05$), but we recognize that p-values alone are not always sufficient in drawing conclusions about relationships or causality (Wasserstein and Lazar, 2016).

We tested several variations of multivariate linear regression models using combinations of the following independent variables: data format count, contributor count, platform count, sensor count, spatial coverage count, spatial resolution count, location count, keyword count, and citation year (as a categorical variable). We ran linear regression analysis in Python using the StatsModels version 0.6.1 (Perktold, Seabold, & Taylor, 2014) module for Ordinary Least Squares (OLS). We regressed the $\log_e$ of the download counts on various combinations of the independent variables.

In order to choose independent variables, we took into account the p-values of each variable within the OLS linear regression models, used recursive feature elimination, randomized LASSO, and normalized linear regression (ridge regression) available as part of the Scikit-Learn module.

3. Results

3.1 Decision Trees

Each of our regression tree models used $\log_e$ of download rate as the dependent variable. The regression tree model with the best prediction rate (determined by RMSE, see Table A2 in the appendix) used binary variables for: data formats, keywords, locations, sensors, and platforms (see Figure A1 in the appendix). The RMSE for this model was 0.978 of $\log_e$ download count or 2.66 ($e^{0.978}$) downloads. Items worth noting in this model are the specific metadata elements deemed most important based on Gini measures:

- Keyword: “sea ice pm polar stereo-project”
- Data Format: hdf-eos
- Sensor: ssmi
- Sensor: thir
- Keyword: “agdc-project”
- Data Format: png
- Location: Australia/New Zealand
- Location: Oklahoma
- Platform: satellites

However, these variables are discipline specific. We desired a model that used general metadata elements that could be of interest to other disciplines. We created a regression tree using the following independent variables: citation year, data format count, spatial coverage count, sensor count, update frequency, and keyword count. The tree was then pruned using the variables determined to be the best predictors of $\log_e$ download count by the cross-validated error rate in RPart's Complexity Parameter Table (Milborrow, 2014). The pruned model used the following variables: citation year, data format count, keyword count, and sensor count (see Figure 5). This regression tree model yielded a prediction RMSE of 1.296 $\log_e$ download counts or 3.655 ($e^{1.296}$) downloads.

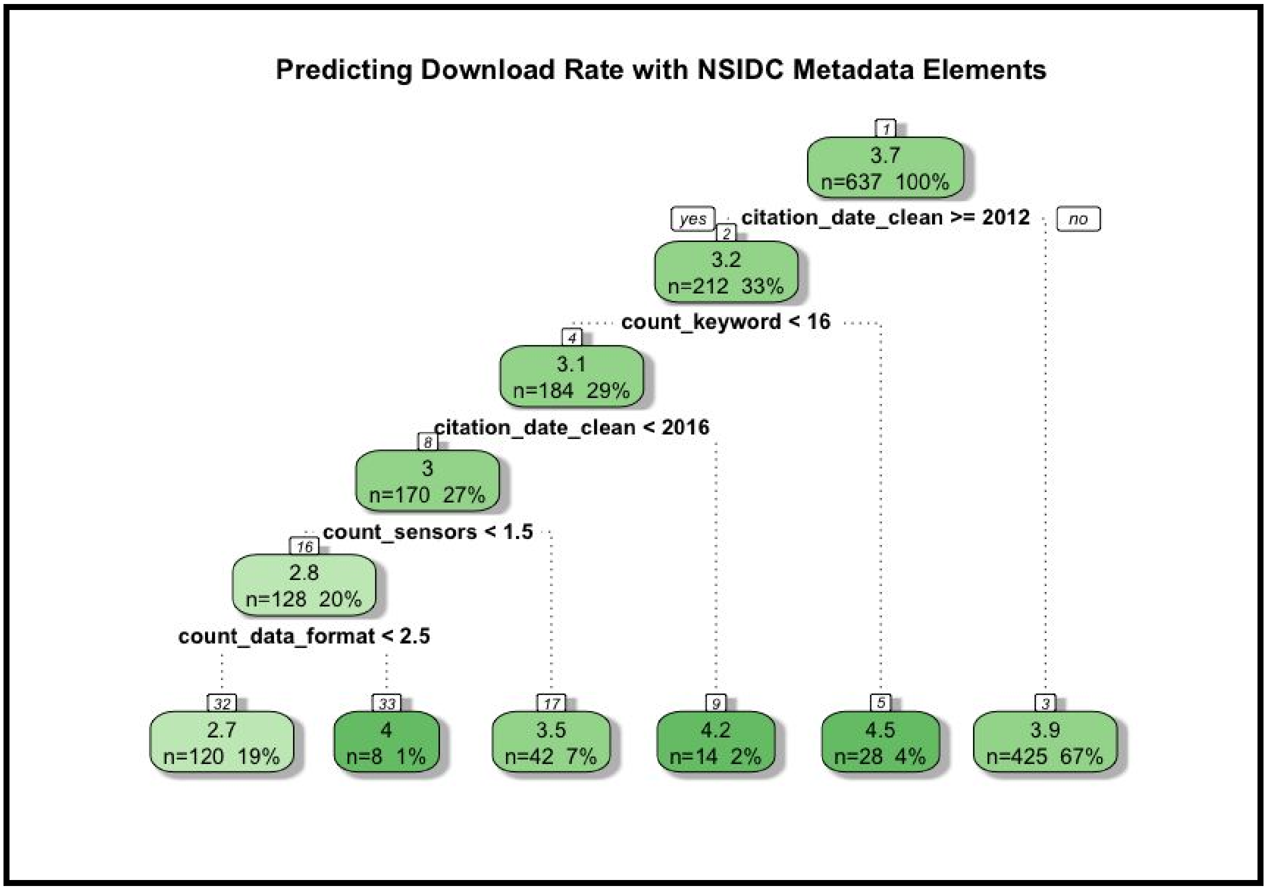


Figure 5: Decision tree using training data with $\log_e$ RMSE (1.296 $\log_e$). Unlabeled numeric in each box represents the mean $\log_e$ of download rates for the given sample (n) of the data. Created with R package RPart version 4.1.10 (Milborrow, 2014) using colors from RColorBrewer version 1.1.2 (Neuwirth, 2011).

3.2 Linear Regression

As with regression trees, our linear models used the $\log_e$ of the number of downloads as the dependent variable. We began by using all eight of the integer features as the independent variables (data format count, contributor count, platform count, sensor count, spatial coverage count, spatial resolution count, location count, keyword count). Six of the eight variables had statistically significant p-values (< 0.05), indicating positive correlation with the dependent variable. The sensor count and spatial resolution count did not correlate with $\log_e$ of the download rate. Removing these two variables resulted in a model with six statistically-significant variables and an adjusted R^2 value of 0.883. We tested predicted outcomes on the testing dataset which yielded a RMSE of 1.42 for $\log_e$ download rate. (For detailed results of each regression model see Table A3 in the appendix.)

As mentioned in section 2.1, to test the potential effect of regularly updated datasets, we removed all datasets that had an update frequency of daily or yearly. We compared the linear regression results of this data to the linear regression results of the eight integer features discussed above. The R^2 and adjusted R^2 values were the same for both models. The same six independent variables had p-values < 0.05. We were hesitant to remove observations from our dataset and the similarity between the models convinced us to continue using the full dataset for the remainder of the modeling.

We then added in the categorical variable of citation year. This addition caused the p-value for location count to rise above 0.05. After trying various iterations of the variables, we found the linear model with the lowest RMSE (1.22 $\log_e$ downloads or 3.39 downloads) to include citation year, data format count, contributor count, platform count, and keyword count. This is was a slight improvement on our general element regression tree.

4. Discussion

From the results of our regression trees and linear regression models, we have shown that certain general and specific metadata features correlate with download rates of NSIDC datasets. Specifically, the general categories of: year of citation, number of data formats, number of contributors, number of platforms, number of spatial coverage areas, number of locations, and number of keywords are positively correlated with download rates. Specific metadata elements such as the keyword “sea ice pm polar stereo-project,” the data format “hdf-eos,” and the sensor type “ssmi,” ranked highly on the Gini criterion of feature importance in regression tree models.

Our strongest prediction model was a regression tree using discipline-specific binary variables with an RMSE of 2.66 downloads ($e^{0.978}$). Our strongest prediction model with more general metadata elements was a linear regression model with and RMSE of 3.39 downloads ($e^{1.22}$). However, we cannot infer causality from our modeling. While certain metadata characteristics are correlated with download rates, it is entirely possible that high interest in a dataset leads to more detailed metadata. Future research could include an analysis of the NSIDC’s data ingest process and levels of service to determine whether extant metadata is changed based on the level of interest in certain datasets.

While we only studied the NSIDC data repository, and our results should not blindly be generalized to other data repositories, especially those outside the earth sciences, they do indicate that higher counts of certain metadata elements are associated with higher download rates. This suggests that larger research projects that output a wider range of data are correlated with higher download rates.

Further research could address causality via experimentation, other data repositories, network analysis of linked datasets, correlation between download rates and reuse rates, and additional machine learning algorithms. As mentioned earlier in the section on definitions of reuse, additional research could also use citation linking and citation counts rather than download rates as our dependent variable.

5. Conclusion

Our recommendation, based on the results of this study, is for both dataset producers and data managers to ensure inclusion of comprehensive metadata, with particular emphasis on keywords, data formats, data collection locations, and then, of course, any other fields of high relevance to the discipline. The Inter-university Consortium for Political and Social Research includes the following statement in their Guide to Social Science Data Preparation and Archiving: “as metadata are often the only form of communication between the secondary analyst and the data producer, good descriptive metadata are essential for effective data use” (p. 11, 2012) and we would add that such metadata are essential for data discovery and reuse.

Acknowledgement

We are indebted to the National Snow and Ice Data Center for providing us with download statistics and dataset IDs for datasets residing in their repository. This research would not have been possible without their important contribution.

References

Belter, C. W. (2014). Measuring the value of research data: a citation analysis of oceanographic data sets. *PloS One*, *9*(3), e92590.

Gelman, A., & Hill, Jennifer. (2007). *Data analysis using regression and multilevel/hierarchical models.* (Series: Analytical methods for social research). Cambridge ; New York: Cambridge University Press.

Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science (New York, N.Y.), 331*(6018), 649. http://doi.org/10.1126/science.1203354

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A Transformed Scientific Method. In *The fourth paradigm data-intensive scientific discovery* (pp. Xvii-Xxxi). Redmond, WA: Microsoft Corporation. Retrieved May 28, 2016 from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI.

Kim, Y., Zhang, P., Bellini, J., Crowston, K., Driscoll, C., Morarescu, P., Qin, J. & Stanton, J. (2013). *Institutional and Individual Influences on Scientists' Data Sharing Behaviors,* ProQuest Dissertations and Theses. doi: 10.1002/meet.14505001093

McKinney, W. ((2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56.

Milborrow S (2014). rpart.plot: Plot rpart models. An enhanced version of plot.rpart. R package version 4.1.10, URL http://CRAN.R-project.org/package=rpart.plot.

National Snow and Ice Data Center. (n.d.). Retrieved May 29, 2016, from https://nsidc.org/about/policies#accordion-3

National Snow and Ice Data Center: Levels of Service (Publication). (2009, July 31). Retrieved May 30, 2016, from National Snow and Ice Data Center website: https://nsidc.org/sites/nsidc.org/files/files/NSIDCLevelsOfService-V2_0a(2).pdf

Neuwirth, E. (2011). RColorBrewer: ColorBrewer palettes. R package version 1.1.2, URL http://CRAN.R-project.org/package=RColorBrewer.

Overpeck, J., Meehl, G., Bony, S., & Easterling, D. (2011). Climate data challenges in the 21st century. *Science (New York, N.Y.), 331*(6018), 700-2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research, 12*, 2825-2830.

Perktold, J., Seabold, S., & Taylor, J. (2014, December 2). StatsModel (Version 0.6.1) [Computer software]. Retrieved from http://statsmodels.sourceforge.net/stable/index.html

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Python Software Foundation. (2015, December 5). Python (Version 2.7.11). Retrieved from http://www.python.org

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 62-71.

Reitz, K., Benfield, C., & Cordasco, I. (2016, April 29). Requests (Version 2.10.0) [Computer software]. Retrieved from http://docs.python-requests.org/en/master/

Richardson, L. (2015, September 28). BeautifulSoup (Version 4.4.1) [Computer software]. Retrieved from http://www.crummy.com/software/BeautifulSoup/bs4/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE, 6*(6), PLoS ONE, 2011, Vol.6(6). DOI**:**10.1371/journal.pone.0021101

Tenopir, C., Dalton, E., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PloS One, 10*(8), E0134826. DOI:10.1371/journal.pone.0134826

Uhlir, P.F., (2010). Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data. *Data Science Journal*, 9, ES1–ES5. DOI: http://doi.org/10.2481/dsj.Essay-001-Uhlir

van der Walt, S., Colbert, C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, **13**, 22-30. DOI:10.1109/MCSE.2011.37

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*. DOI:10.1080/00031305.2016.1154108

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1-2), 5–16. DOI:10.1007/s00799-007-0015-8

Appendix

Variable	Description	Type of values	Scale
dataset_id	Unique identifier assigned to dataset by NSIDC	Discrete	Nominal
unique_users_ip	The number of downloads from unique IP addresses associated with each of the dataset IDs	Continuous	Ordinal
scrape_date	Date of scraping associated metadata provided on the NSIDC's public website	Discrete	Nominal
scrape_time	Time of scraping associated metadata provided on the NSIDC's public website	Discrete	Nominal
version_clean	Version number of dataset	Continuous	Ordinal
title_original	Original title of dataset	Discrete	Nominal
doi_address_clean	Digital object identifier for dataset	Discrete	Nominal
citation_date_clean	Year included in citation for dataset	Discrete	Ordinal
count_data_format	Number of data formats available for dataset	Continuous	Ordinal
data_format_string	List of data formats available within the dataset	Discrete	Nominal and Binary
contributors_clean	List of contributors for each dataset in "First name Last name" format.	Discrete	Nominal
contributor_list	List of contributors for each dataset in "First letter of the first name - Last name" format.	Discrete	Nominal and Binary
contributor_last_names	List of last names of contributors for each dataset.	Discrete	Nominal and Binary
count_contributors	Number of contributors for each dataset.	Continuous	Ordinal
spatial_coverage_clean	Spatial coverage for each dataset	Discrete	Nominal

count_spatial_coverage	Number of locations (spatial coverage)	Continuous	Ordinal
spatial_resolution_clean	Spatial resolution descriptions	Discrete	Nominal
count_spatial_resolution	Number of spatial resolutions for each dataset	Continuous	Ordinal
platforms_clean	List of platforms for each dataset	Discrete	Nominal and Binary
count_platforms	Number of platforms for each dataset	Continuous	Ordinal
sensors_clean	List of sensors for each dataset.	Discrete	Nominal and Binary
count_sensors	Number of sensors for each dataset	Continuous	Ordinal
update_frequency	How often each dataset is updated	Discrete	Nominal and Binary
location_clean	List of location names for each dataset	Discrete	Nominal and Binary
count_locations	Number of location (names) for each dataset	Continuous	Ordinal
keyword_clean	List of keywords for each dataset	Discrete	Nominal and Binary
count_keyword	Number of keywords for each dataset	Continuous	Ordinal
last_updated	The latest date when the dataset was updated in mm/dd/yyyy format.	Discrete	Ordinal
count_page_ref	Number of links to web pages in the "See also" section of the dataset description.	Continuous	Ordinal
count_dataset_ref_out	Number of links to other datasets in the "See also" section of the dataset description.	Continuous	Ordinal
dataset_ref_out	Id-s of datasets in the "See also" section of the dataset description.	Discrete	Nominal
count_dataset_ref_in	Number of datasets that have references to the dataset.	Continuous	Ordinal

dataset_ref_in	Id-s of datasets that have references to this dataset in their "See also" sections.	Discrete	Nominal

Table A1: Full list of variables, data types, and scales included in our dataset.

Software/ Library of Model Generation	Additional Parameters	Initial Independent Variables	Variables Chosen by Model Parameters	RMSE of Log_e Predicted Test Download Rates vs. Log_e Actual Test Download Rates
R/RPart	Number of Splits = lowest cross-validated error rate, method=anova	count_data_format, citation_date_clean, count_spatial_coverage, count_sensors, update_frequency, count_keyword	citation_date_clean, count_data_format, count_keyword, and count_sensors.	1.296
Python/SKLearn	Max Nodes = 10	All Dummy Variables from Data Formats, Keywords, Locations, Sensors, and Platforms	“sea ice pm polar stereo-project,” “hdf-eos,” “ssmi,” “thir,” “agdc-project,” “png,” “australia/new zealand,” “oklahoma,” “satellites”	0.978
Python/SKLearn	Max Nodes = 10	All Sensors	“ssm/i,” “modis,” “dslr,” “thir,” “amsu-a,” “smmr,” “ssmis,” “pals,” “amsr-e”	1.021
Python/SKLearn	Max Nodes = 10	All Keywords	“sea ice pm polar stereo-project,” “agdc-project,” “ease-grid-project,” “sea ice,” “numerical weather prediction,” “glacier fluctuation,” “modis-project,” “smex,” “smap validation cl07-project”	1.143
Python/SKLearn	Max Nodes = 10	All Data Formats	png, hdf-eos, esri-shapefile, hdf, binary, kml, geotiff, microsoft-excel	1.189
Python/SKLearn	Max Nodes = 10	All Locations	Baltic Sea, Oklahoma, New Zealand, Georgia, Bering Sea, Antarctica, Arctic, Mexico, Maryland	1.19

Table A2: Results of regression decision tree models.

Decision Tree Using Binary Variables

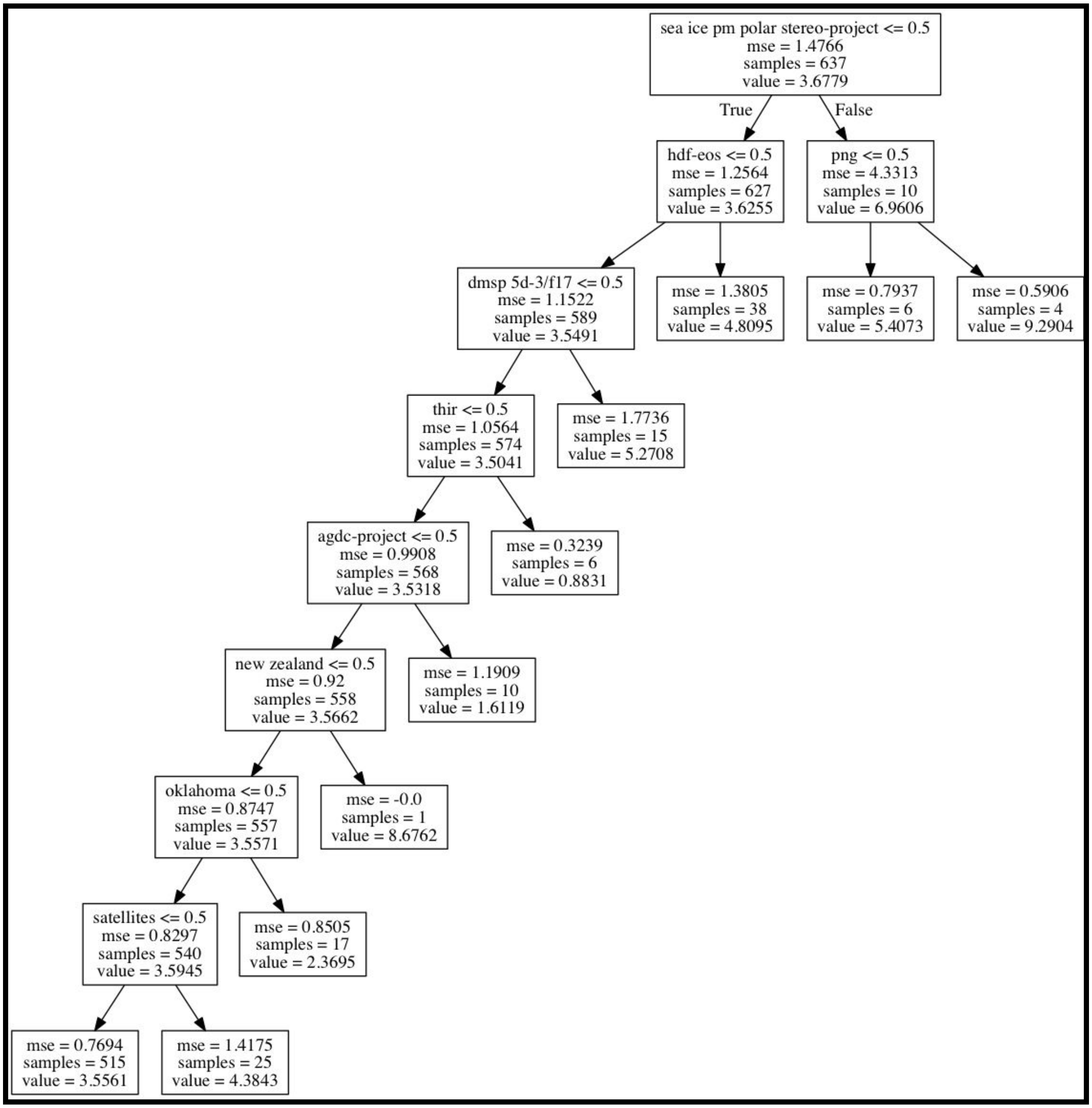


Figure A1: Decision tree using dependent variable: $\log_e$ of download and independent variables: binary (dummy) of data formats, keywords, locations, sensors, platforms. Tree built with Python and Scikit-Learn.

Independent Variables Used in OLS Linear Regression Model	Coefficient Value for Each Variable	P-Values for Each Variable * Indicates $p < 0.05$	RMSE of Log$_e$ Predicted Test Download Rates vs. Log$_e$ Actual Test Download Rates	Model R-squared	Model Adjusted R-squared
count_data_format	0.3693	4.242556e-10*	1.4201	0.884	0.882
count_contributors	0.2391	8.105154e-14*			
count_platforms	0.1894	2.392165e-07*			
count_sensors	0.0101	7.445888e-01			
count_spatial_coverage	1.0941	2.212030e-24*			
count_spatial_resolution	0.0269	7.818940e-01			
count_locations	0.1870	2.990326e-08*			
count_keyword	0.0350	6.767895e-07*			
count_data_format	0.3693	4.098764e-10*	1.4204	0.884	0.882
count_contributors	0.2401	4.607525e-14*			
count_platforms	0.1947	3.230500e-09*			
count_spatial_coverage	1.0918	2.002456e-24*			
count_spatial_resolution	0.0245	8.001529e-01			
count_locations	0.1878	2.345658e-08*			
count_keyword	0.0356	1.400702e-07*			
count_data_format	0.3710	2.517681e-10*	1.4208	0.884	0.883
count_contributors	0.2417	9.482735e-15*			
count_platforms	0.1962	1.349885e-09*			
count_spatial_coverage	1.0958	4.114819e-25*			
count_locations	0.1864	1.875876e-08*			
count_keyword	0.0356	1.410781e-07*			
count_data_format	0.1754	4.200681e-04*	1.222	0.311	0.277
count_contributors	0.0740	7.935439e-03*			
count_platforms	0.2063	5.676988e-15*			
count_keyword	0.0175	1.875624e-03*			
citation_date:1983	2.9411	5.180690e-05*			
citation_date:1984	-8.367e-14	1.079456e-07*			
citation_date:1988	4.3765	1.921379e-05*			
citation_date:1991	4.7402	3.428281e-06*			
citation_date:1992	3.6468	1.091676e-09*			
citation_date:1994	4.3283	2.675702e-05*			
citation_date:1995	3.7574	1.587840e-23*			
citation_date:1996	4.7225	2.254767e-10*			
citation_date:1997	2.6259	8.294835e-10*			
citation_date:1998	3.1623	6.210551e-56*			
citation_date:1999	3.4871	9.427533e-24*			
citation_date:2000	3.3666	1.309116e-14*			
citation_date:2001	3.6542	3.097338e-19*			

citation_date:2002	3.5891	2.124825e-48*			
citation_date:2003	2.9141	1.932465e-62*			
citation_date:2004	2.8958	1.435540e-48*			
citation_date:2005	2.4589	2.158359e-16*			
citation_date:2006	3.4886	2.790683e-46*			
citation_date:2007	2.8078	5.042582e-29*			
citation_date:2008	2.6982	1.103739e-28*			
citation_date:2009	2.6167	7.260191e-44*			
citation_date:2010	2.7180	1.026754e-36*			
citation_date:2011	2.5990	1.066909e-27*			
citation_date:2012	2.9447	3.062377e-22*			
citation_date:2013	1.9419	9.813542e-27*			
citation_date:2014	] 2.4009	4.130774e-38*			
citation_date:2015	1.9928	6.921112e-32*			
citation_date:2016	2.8574	2.768895e-22*			

Table A3: Results of multivariate linear regression models using Python (version 2.7.11) and StatsModel (version 0.6.1) module for Ordinary Least Squares regression.