THE INTERNATIONAL SOCIETY OF
PRECISION AGRICULTURE PRESENTS THE
13th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE
July 31-August 4, 2016 • St. Louis, Missouri USA

Window-based regression analysis of field data

Anne M. Denton, Harshada Chavan, David W. Franzen, and John F. Nowatzki

Department of Computer Science, North Dakota State University, Fargo, ND

Department of Computer Science, University of Minnesota, Minneapolis, MN

Department of Soil Science, North Dakota State University, Fargo, ND

Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND

A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA

Abstract.

High-resolution satellite and areal imagery enables multi-scale analysis that has previously been impossible. We consider the task of localized linear regression and show that window-based techniques can return results at different length scales with very high efficiency. The ability of inspecting multiple length scales is important for distinguishing factors that vary over different length scales. For example, variations in fertilization are expected to occur on shorter length scales than changes in soil type. We demonstrate the effectiveness of our approach for a small agriculturally relevant use case, in which regression lines are calculated for the dependency of yield on the Normalized Difference Vegetation Index, NDVI. This use case is relevant towards the In Season Estimation of Yield, INSEY. Conventionally, yield vs. NDVI dependencies are established based on data collected for test plots. However, the results from tests plots may not be representative of the growing conditions in a particular production field. On the other hand, when production-field data are used, dependencies on soil types and other factors may interfere with the fertilization-dependency that is of interest. Our approach promises to allow distinguishing such factors, provided they result in variations on different scales. We compare our technique with Geographically Weighted Regression, GWR. Even a single application of GWR takes over one hour for 10,000 data points, while our own approach completes in under one minute while, at the same time, returning multiple maps, each corresponding to a different resolution.

Keywords. *Window-based analysis, Regression, High-resolution images*

Introduction

Remotely sensed data are becoming available at rapidly increasing resolutions. Yet, the techniques that are available for their analysis were developed when the number of data points per field was in the hundreds not millions. For over four decades, the Landsat missions have provided imagery with a 30m resolution free of charge, but higher resolution is quickly becoming available through private satellite providers and aerial imagery, including from unmanned air vehicles. Traditional statistics approaches for localized regression, such as Geographically Weighted Regression, GWR, (Fotheringham et al. 2003), were designed for a few hundred data points, millions, and at 1m resolution a 160 acre agricultural field is covered by almost one million data points.

This volume of data allows asking questions that have not previously been asked. In particular, the resolution-dependence of regression results can be studied over a wide range of length scales. The growing conditions are expected to depend on variables that change over several different length scales, with fertilizer dependence varying over the shortest length scale, hydrological parameters over intermediate length scales, and soil types over length scales not much smaller than field size. However, even a single application of GWR with 10,000 data points takes of the order of one hour, so exploring resolution dependence for high resolution data is out of the question with traditional techniques. In our work, we limit ourselves to linear regression, for which it is possible to double the length scale of aggregation by aggregating data from the previous iteration further. Figure 1 shows the concept. Each of the larger window has four sub-windows that acted as the main windows in the previous iterations. Using this approach, the complexity per pixel in the image is log *w* since the region over which aggregation is accomplished doubles in each iteration.

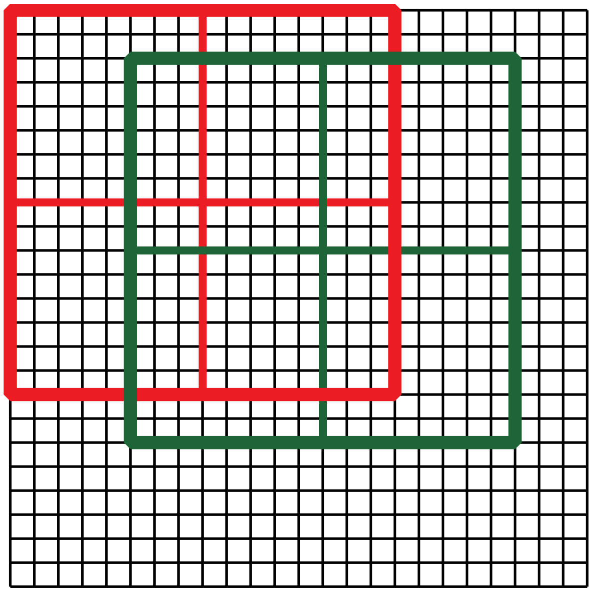

Figure 1: Schematic for sliding window extraction

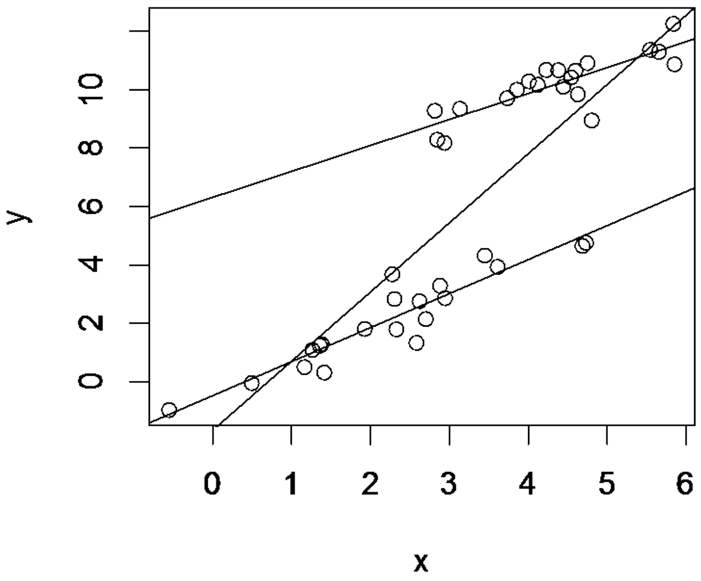

Figure 2: Schematic showing the problem of fitting multiple subpopulations together

We consider the task of deriving linear regression results that can be used for in-season fertilization. The In Season Estimate of Yield is also abbreviated as INSEY (Solie 2000). Conventionally, the dependence of yield on NDVI is derived by soil scientists based on the results for experimental plots (Franzen 2011). There are several limitations to this approach. In particular, it is known that the dependencies vary for different types of soil, and that soil maps are often inaccurate. Being able to derive the yield vs. NDVI relationship from the actual production field data would be valuable. The direct approach of applying regression to field-level data suffers from the lack of statistical rigor that is typically the result of an analysis that is based on existing data.

Applying data mining techniques in this environment has to be done with care. Traditionally the problem of in-field variations is addressed through the statistics of experimental design. For a complete production field, experimental design requirements cannot be met, and yield may vary for any number of extraneous reasons. Regression slopes that are derived for a complete field may not be accurate for any portion of the field, and even localized regression results may only compensate for one of the sources of variations, such as fertilization, hydrology, or soil type. Figure 2 shows the principle problem. Let us assume that the points in the upper portion and lower portion of the figure are from two different regions in a field. Taken by themselves, each set of points can be represented well through a regression line. However, when all points are considered together, the resulting regression line is not an appropriate representation of either of its subsets. The slope that would be deduced from the combination of regions would not provide a farmer with correct recommendations, and localized regression done at one length scale may not capture the most reliable dependency. Making multi-scalar analysis performant enough for this problem is a central step towards a true data mining solution to agricultural data mining.

Related Approaches

Many other data mining algorithms have been developed specifically for geospatial data (Mennis & Guo 2009). Arguably the most closely related technique in comparison to the presented approach is geographically weighted regression, proposed by Fotheringham et al. (Fotheringham et al. 2003). Another regional regression approach that was introduced in (Celepcikay & Eick 2009) uses regional dependencies between the dependent and independent attributes. While much of the work addresses point data, possibly involving time (Zhou et al. 2011), some algorithms have been designed specifically for remotely sensed raster data (Zhang et al. 2008). The concept of windows is often used for smoothing (Anselin & McCann 2009) but has also been used for defining spatial co-location patterns (Shekhar & Huang 2001). Data of multiple granularities has also been used in climate models (Subbian & Banerjee 2013). However, the concept of analyzing geospatial raster data on multiple length scales is not part of the standard geospatial data mining toolbox.

The agriculture domain has many applications for data mining and (Mucherino et al. 2009) summarizes some work such as *k*-means, *k*-NN classification, SVM, etc. in agriculture. The geospatial nature of agriculturally relevant satellite image data affects standard statistics and data mining tasks (Moran & Bui 2002). Data mining techniques are often used to study soil characteristics. For example, (Meyer et al.) apply the *k*-means algorithm to finding clusters of soils and plants. Much work has also been done on predicting yield (Russ 2009).

Concepts

The proposed algorithm has three steps:

1. Iterative sliding-window-based extraction of slopes
2. Histogram-based evaluation of appropriate window size and slope cutoffs
3. Geospatial rendering of slope clusters

Note that the window-based extraction creates some spatial continuity by design, but there is no guarantee that slopes will result in patterns beyond that. The histogram-based evaluation allows identifying meaningful breaking points in the data itself. Sliding-window-based techniques use all sub-windows of a given image. For a window-size of *w*, and an image of width *n* and height *m*, there are *(n-w+1)(m-w+1)* sub-windows to consider. By default, any of those windows would have to be parsed, resulting in an overall complexity that is proportional to w^2. One might consider the alternative approach of parsing the overall image once and adding each pixel to all of the windows that contain it. However, since this approach requires adding the pixel to w^2 windows, it does not fundamentally solve the problem of having the complexity depend quadratically on the window-size. Fortunately, the slope of a linear regression line only depends on additive quantities. Not all quantities that can be calculated over a window satisfy the additive property. The median is a common example of a quantity for which such an approach does not work. The additive property allows breaking down each window into four sub-windows, see Figure 1. The figure highlights two windows, one red and one green and, for each of them, it also highlights four sub-windows that are assumed to have been computed during the previous iteration.

Experiments

Data Processing

The field-level data is from the commercial satellite, RapidEye, and we got the data through SatShot (SatShot 2015). The resolution of these images is 5m. The specific images in the evaluation are from the year 2012 for a wheat field in Stutsman County, ND, and were taken in the month of June. The size of the field under consideration was about 16 hectares. The large-scale evaluations are done on satellite images from the Landsat satellite (Landsat 2015) which images the entire earth every 16 days and collects data from nine spectral bands, of which we use the red and near infrared bands. These images are in the raster format, with a resolution of 30 m.

Large-scale Example

In our first experiment we apply the algorithm to the near infrared and red bands of a Landsat image of 1000 by 1000 pixels. The result of an aggregation to window size 4 can be seen in Figure 3. The computation time for this image was under 1 min. For comparison purposes, we tested GWR on the same image, but we had to reduce the size to 100 by 100 pixels, and even at that size the program took more than one hour to complete for a single resolution. There is no doubt that the quality of results is higher for GWR, since sophisticated kernel methods are used instead of basic linear regression. The scope of both approaches is very different in that we aim to provide a highly performing algorithm that returns basic linear regression results over sliding windows on multiple length scales, while for GWR the full focus is on the quality of the regression.

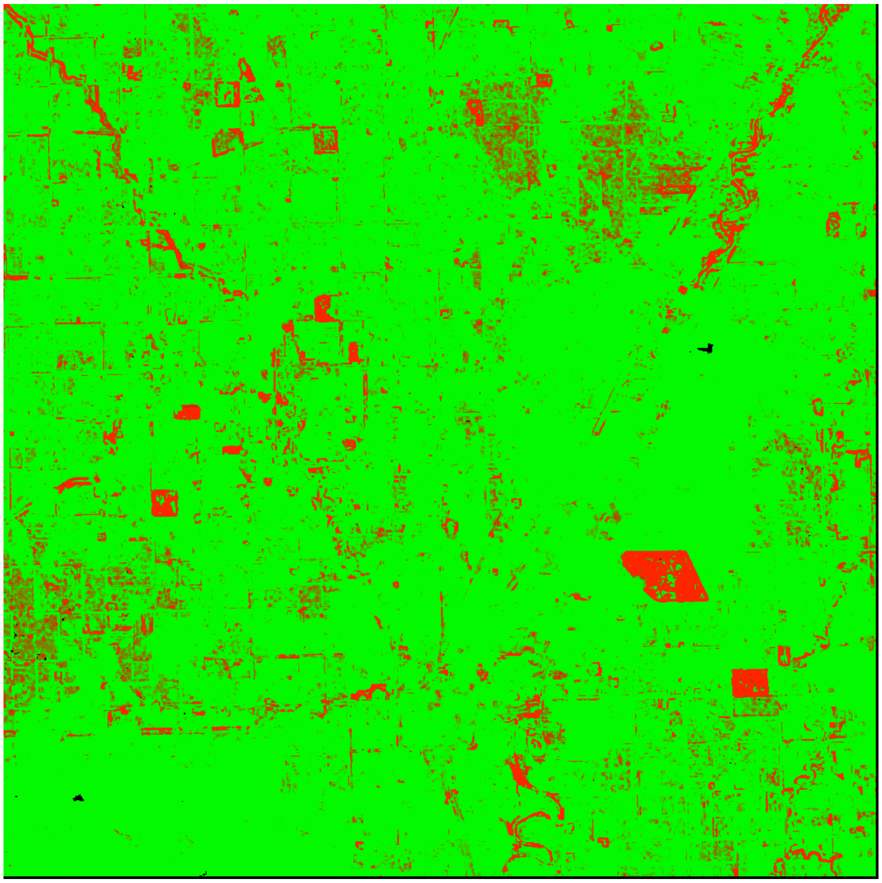

Figure 3: Slopes over *w*=4 windows of the regression slope of the near infrared band vs. the red band of a Landsat Image. False color is used with green corresponding to positive slopes and red to negative ones.

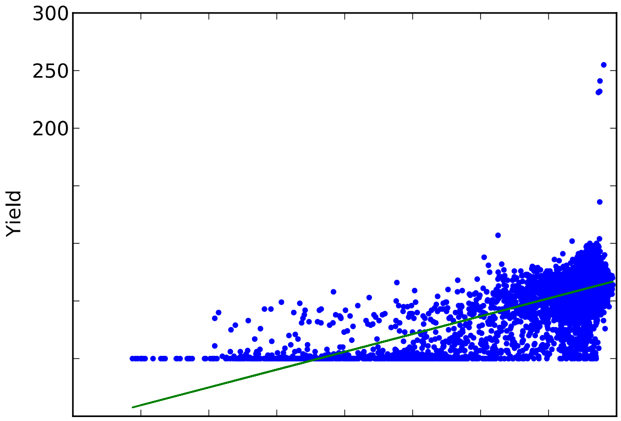

Figure 4: Scatter plot of yield vs. NDVI for a complete field (blue) together with linear regression line (green)

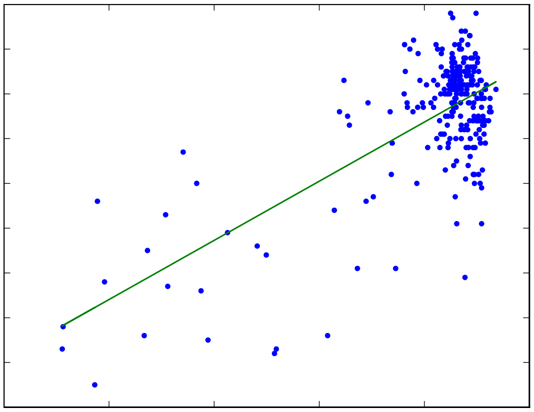

Figure 5: Scatter plot of yield vs. NDVI for a window of size 16 x 16 (blue) together with linear regression line (green)

Results for Effectiveness

To establish the value of the multi-scalar analysis we used a small field for which we had expert opinion available. The Figure 4 shows the overall linear regression line for yield against NDVI collectively for the whole field. Windows cover a smaller area, and can show a more coherent distribution, as can be seen in Fig. 5. The noisy distribution is to be expected for satellite imagery of fields, but there is a clear increase of yield with NDVI that can be captured in the value of the slope. The noisiness of the data is such that going beyond linear regression is not expected to produce reliable results.

Figure 6 shows the distribution of slopes over all sliding windows in the field for a window size of 16 x 16. It can be seen that there is some structure. The bulk of slopes is between about 80 and 250, although there is also a substantial maximum at somewhat negative values. The maximum is slightly higher than the overall slope of 155, and there is no reason to assume that both should be identical. We use 80 as a cutoff for displaying the distribution of slopes across the field. Figure 7 shows the distribution of windows with slopes above and below 80 respectively.

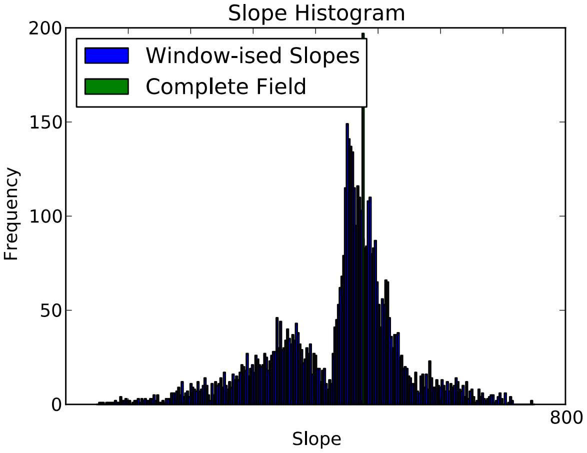

Figure 6: Histogram for window size 16 x 16

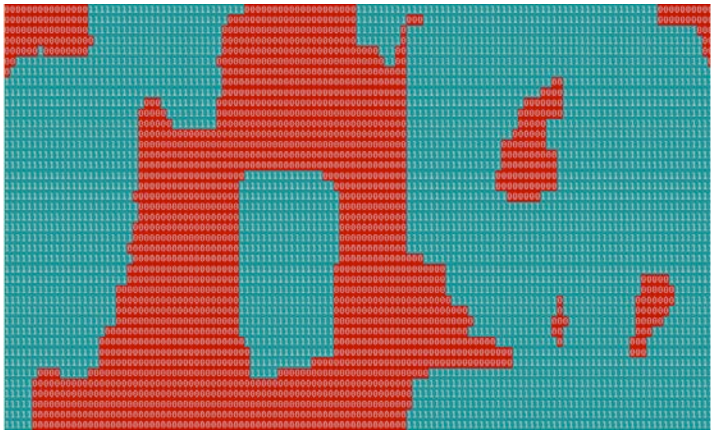

Figure 7: Distribution of slopes below vs. above 80 across the field

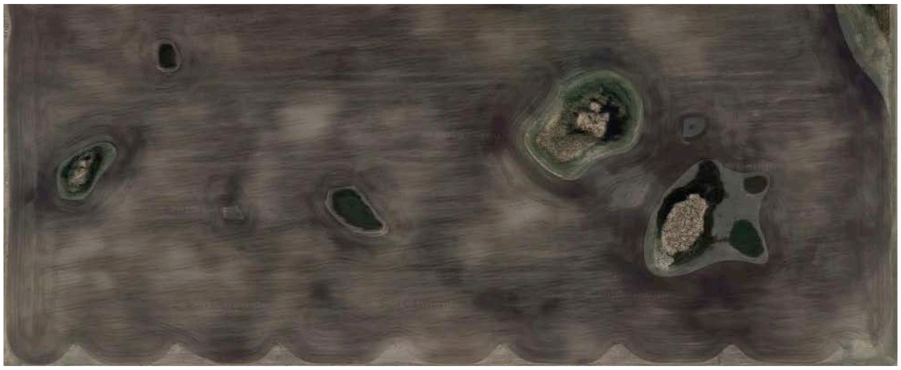

Figure 8: Image of the field using the "Satellite" selection in Google Maps

Windows with small values of the slopes are represented by small red rectangles that are labeled as 0, and windows with large slopes are labeled as 1. It can be seen that there are large sections that show coherent slopes, and that some of those sections are substantially larger than the window size of 16 x 16.

Comparing with areal imagery of the field from Google Maps (collected using "Satellite" setting), in Figure 8 shows that some clusters seen in the spatial distribution of slopes have matching features in the imagery. In particular, the two areas that stand out towards the right of the Google Maps image have corresponding clusters of small slopes. Notice that there is some level of distortion, and that the edges are not fully represented, because only *(n-w+1)* x *(m-w+1)* windows can be constructed for a *n* x *m* image.

The distinction into areas with low slopes for the yield vs. NDVI dependence makes sense from an agricultural perspective. The areas with unusual dependence of yield on NDVI are likely to be areas in which something went wrong or not entirely as planned during the growing season. We tested whether combinations of attributes, in particular mean yield, mean NDVI, slope, and error together provided a crisper definition of clusters but that was not the case. When combinations of attributes were used, the cluster centers differed so little in any one of the attributes that an interpretation would have been inconclusive. Using slope alone as attribute removes ambiguity over what may be represented. A user may then have different questions of the data but those can be answered through querying the data.

When going to smaller window sizes, noise makes slopes increasingly unpredictable. Figure 9 shows a histogram of slopes for windows of 4 x 4 pixels. It can be seen that slopes can be as low as -1000 and as high as 1000. The figure also shows that there is not much structure to the histogram, i.e. slopes are most commonly around 0 and the prevalence gradually decreases for values further away from 0, but there are no clear minima other than through random variations. When applying smoothing to the histogram, any local minima disappear.

Figure 10 shows the distribution of positive and negative slopes spatially across the field. It is not surprising that there is very little structure to the spatial distribution as well. Neighboring windows do have some degree of correlation but some correlation is expected because of the overlap in pixels alone. Two neighboring 4 x 4 windows share 12 of the 16 pixels in the window, enough to result in some correlation even for random data. It is also interesting to observe that, at this small window size, there is no clear maximum anywhere near the overall slope of the field, although there are more positive slopes than negative ones.

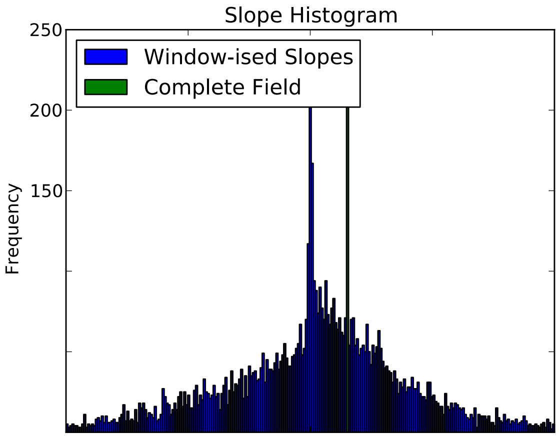

Figure 9: Histogram for window size 4 x 4

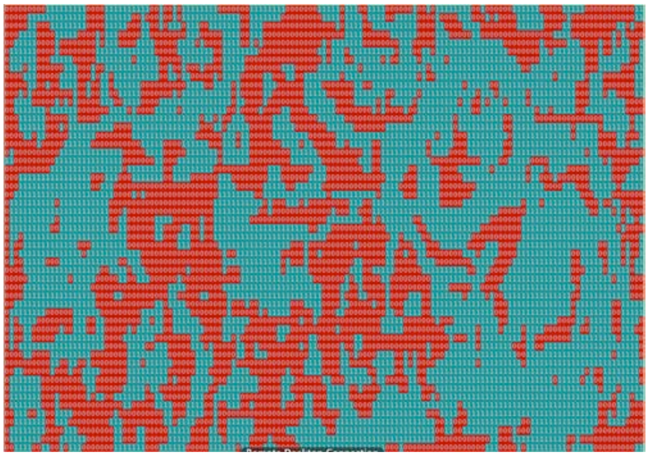

Figure 10: Distribution of positive and negative slopes across the field

Finally, we used a window size of 32 x 32, at which size the window takes up a substantial portion of the overall image, which is 57 pixels in its smaller dimension. The histogram in Figure 11 shows that slopes are now mostly limited to a range of 100 to 200 and show two maxima that are not longer as cleanly separated as for the 16 x 16 window size. The overall slope for the field of 155 is close to the minimum, which is at about 170. We use 170 as cutoff for the geospatial analysis. If we kept using 80, as for 16 x 16 windows, we would now see very few windows that are red.

Figure 12 shows that the region of high slope towards the left side of the image, which was visible in Figure 7, is still noticeable, and the region with low slope to the right of it is as well. However, the distinctive features from the Google Maps image can no longer be recognized, and couldn't really be since their dimensions are substantially smaller than the window size in this analysis. Overall we found that the window size 16 x 16 is most faithful to what would be expected from other considerations, and also had the most clearly identifiable maxima in the histogram representation.

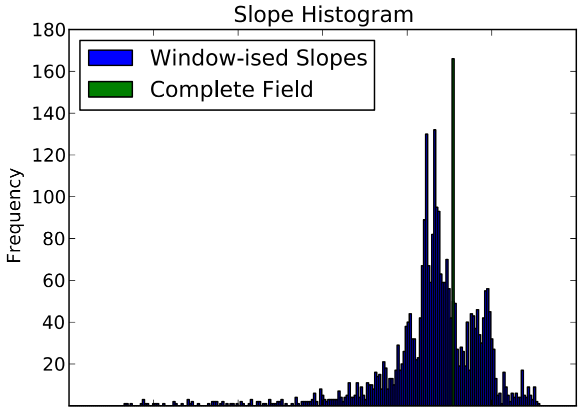

Figure 11: Histogram for window size 32 x 32

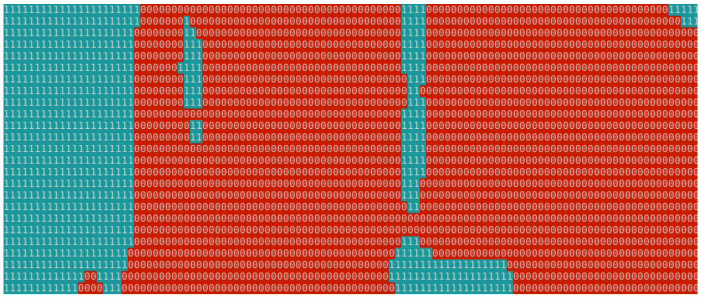

Figure 12: Distribution of slopes below vs. above 17 across the field

We also considered whether the root mean square error, RMSE, may provide a more direct measure of appropriate window size choice than the analysis of the histograms. Table 1 shows the average RMSE values for the window sizes we considered, and there is no clear structure. Smaller window sizes are expected to have a lower RMSE value even for random data, because the RMSE is calculated with regard to the best fit line, which can fit more closely when fewer data points are considered. The decrease of the RMSE for small windows is more than would be expected based on this observation, which is supported by the histogram-based analysis. However, it does not suggest a clear favorite among the window sizes. For this reason, we used the histogram-based analysis as main criterion for identifying the most appropriate window size.

Table 1: Root Mean Square Error for Different Window Sizes

Window Side Length	RMSE
4	8.49
8	11.00
16	13.13
32	14.54

Results for Efficiency

For the runtime evaluation, we used Landsat data to allow us to go up to higher image and window sizes. Figure 13 shows the runtime as a function of window size. It can be seen that the scaling is logarithmic in the side length of the window for the iterative algorithm and quadratic for the default algorithm. In the iterative algorithm every iteration corresponds to twice the window size of the previous side. For the largest window size of *w*=128 there are almost three orders of magnitude difference between both algorithms. The default algorithm requires 1024 times as many operations on window pixels for *w*=128 as for *w*=4. In practice, the runtime increase is somewhat smaller, due to the overhead of reading the image and writing out results, but the quadratic scaling is clearly visible from the figure. For the iterative algorithm, the run for window size of *w*=128 requires 5 times as many iterations as for *w*=4, resulting in an overall runtime increase by about a factor of 3. Figure 14 shows the runtime as a function of image size for a window size of 16 x 16.

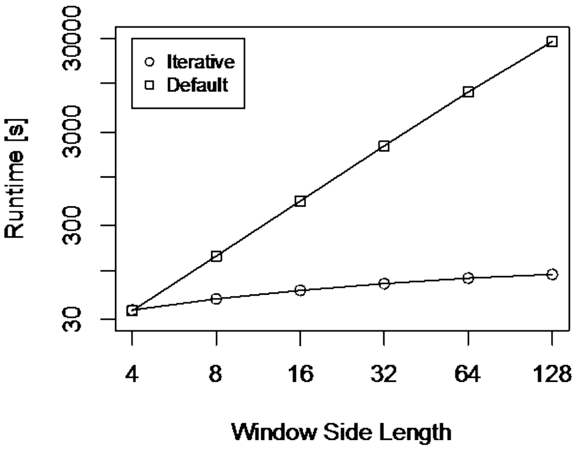

Figure 13: Runtime-dependence on window size

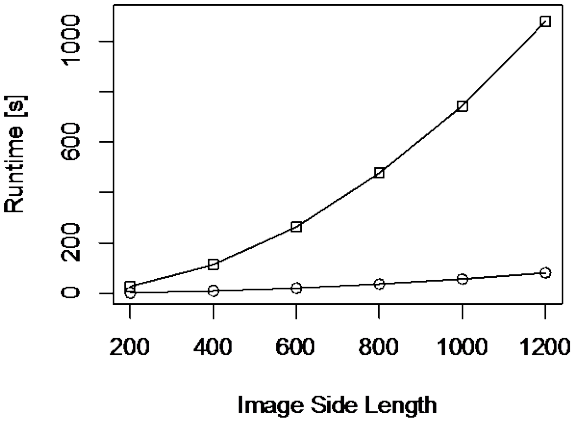

Figure 14: Runtime dependence on image size

It can be seen the runtime for both the iterative algorithm and the default implementation show an increase in runtime that is quadratic in the side length of the image, corresponding to a linear increase in the number of pixels in the image. As expected, both algorithms show the same scaling as a function of image size, but the runtime is about one order of magnitude faster larger for the iterative algorithm than the default implementation.

Conclusions

We have shown the promise and computational feasibility of a multi-scale sliding-window-based analysis of agricultural data. Computational efficiency was gained through an aggregation approach that has a logarithmic number of aggregation steps, each of which consists in adding aggregate quantities from four sub-windows. We have also shown that an analysis at each of those levels is valuable since it may not be clear, a priori, which window size is most promising. Histograms of the slopes of regression lines provide feedback on window sizes with meaningful information content. While a small window size does not show much structure in the distribution of slopes, and a large window size may not discriminate different regions well, we were able to identify clear regions in our intermediate window size of 16 x 16 pixels. For the field that is shown in detail, the analysis was almost instantaneous. For an image of 1200 x 1200 pixels and a window size of 16, the analysis could still be done in about a minute. We did not go to larger images because the default analysis was already impractical, taking about 20 min. For larger window sizes, the difference becomes yet much more substantial. For a window size of 128, the default approach is almost three orders of magnitude slower than the iterative approach. This work is only scratching the surface of what can be done using a multi-scalar analysis of agricultural data. Full multi-scalar approaches in contexts such as simulation environments, allow for different resolutions in different areas. Future work will explore the potential of such flexibility in the analysis of agricultural images.

Acknowledgements

This material is based upon work supported by the National Science Foundation through grants PFI-1114363 and IIA-1355466.

References

Anselin, L. & McCann, M. (2009). Opengeoda, open source software for the exploration and visualization of geospatial data. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, (pp. 550–551). ACM.

Celepcikay, O.U. & Eick, C.F. (2009). Regˆ2: a regional regression framework for geo-referenced datasets. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, (pp. 326–335). ACM.

Fotheringham, A.S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

Franzen, D.W., Endres, G., Ashley, et al. (2011). Revising nitrogen recommendations for wheat in response to the need for support of variable-rate nitrogen application. *J. of Agr. Sci. and Tech.*, A1,89–95.

Landsat (2015). United State Geological Survey. Landsat project. Available at http://landsat.usgs.gov/, Accessed: 4-30-2015.

Mennis, J. & Guo, D. (2009). Spatial data mining and geographic knowledge discovery – an introduction. *Computers, Environment and Urban Systems*, **33**(6):403–408.

Meyer, G., Neto, J., Jones, D. et al. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. *Computers and Electronics in Agriculture*, **42**:161–180.

Moran, C.J. & Bui, E.N. (2002). Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science*, **16**(6):533–549.

Mucherino, A., Papajorgji, P. & Pardalos, P. (2009). A survey of data mining techniques applied to agriculture. *Operational Research*, **9**:121–140.

Russ, G. (2009) Data mining of agricultural yield data: A comparison of regression models. In Proceedings of the 9th Industrial Conference on Advances in Data Mining. Applications and Theoretical Aspects, (pp. 24–37).

Satshot (2015). Available at https://www.satshot.com/, Accessed: 4-30-2015.

Shekhar, S. & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In Advances in Spatial and Temporal Databases, (pp. 236–256). Springer.

Solie, J.B., Stone, M.L, Needham, D.E. et al. (2000). In-season N fertilization using an in-season estimate of potential yield. In Proceedings of the 5th International Conference on Precision Agriculture,(pp. 1–8). American Society of Agronomy.

Subbian, K. & Banerjee, A. (2013). Climate multi-model regression using spatial smoothing. In Proceedings of the 2013 SIAM International Conference on Data Mining (pp. 324–332). SDM.

Zhang, Z., Wu, W. & Huang, Y.. Effective spatio-temporal analysis of remote sensing data. In Progress in WWW Research and Development (pp. 584–589).

Zhou, X., Shekhar, S., Mohan, P. et al. (2011). Discovering interesting sub-paths in spatiotemporal datasets: A summary of results. In Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 44–53). ACM.