Statistical challenges overcome in publishing GIS analyses

George Mueller-Warrant, Research Agronomist, USDA-ARS, Corvallis, OR

ESRI-User Conference, San Diego, CA on July 25, 2012

ABSTRACT. GIS data and methods have the potential to answer questions concerning agriculture that were too complex to even consider formulating until quite recently. Unfortunately for researchers trying to publish findings ranging from the reasons for spatial patterns in the distribution of weeds to options for the placement of straw-based biofuel refineries, the publication process is fraught with extra challenges when GIS enters the picture. In my experience, successfully navigating the peer-review publication process includes not just finding the "right" journal, but also translating the arcane language of geo-statistics into terms and concepts that resonate clearly with editors, reviewers, and ultimately readers of your paper. While painful, being forced to be the first one to describe underlying assumptions behind the colorful maps that take only a few (final) keystrokes to produce is not without its rewards - foremost amongst these is your own improved understanding of how the "tools" actually work.

Statistical challenges overcome in publishing GIS analyses –

"A Tale of Three (or More) Papers"

GIS paper #1 – "Remote sensing classification of grass seed cropping practices in western Oregon" by Mueller-Warrant *et al*. 2011. *International Journal of Remote Sensing* 32(9):2451-2480.

GIS paper #2 – "GIS Analysis of Spatial Clustering and Temporal Change in Weeds of Grass Seed Crops" by Mueller-Warrant, Whittaker, and Young. 2008. *Weed Science* 56:647-669.

GIS paper #3 – "Geospatial identification of optimal straw-to-energy conversion sites in the Pacific Northwest" by Mueller-Warrant, Banowetz, and Whittaker. 2010. *Biofuels, Bioproducts, and Biorefineries* 4:385-407.

Common themes in all three sagas include rejections, revisions, resubmissions, and willingness of the journals to only publish their favorite parts of any story. Final versions of the papers also included lengthy sections describing the methods at a level of detail needed to explain GIS and spatial statistics to the novice – it sometimes/often felt like I was writing the text for "Spatial Stats 101" or maybe "399".

General points:

(1) Journals that specialize in "spatial stuff" – remote sensing, geographic information science/systems, spatial statistics – tend to mainly want the latest idea, the newest version, the faster or better algorithm, the paradigm shift. In other words, they are not so interested in having you take someone else's approach and apply it to farms in the Pacific Northwest rather than corn and soybean fields in Iowa. They already have too many papers being submitted and you are competing for limited space.

(2) Journals that specialize in "everything else" – crops, weeds, agronomy, soil science, water quality, or ecology – tend to have their own favorite ways of analyzing data, ways that have little to do with spatial properties or statistics. They will be interested in how you analyze spatial data and connect those results to more traditional concepts, but you will have to do the heavy lifting of explaining the concepts to the editors and reviewers. And most importantly, you will have to identify the hypotheses you are testing, and probably define them both in terms of GIS/spatial statistics' concepts and in language the reviewers are more familiar with. Getting to that point will force you to clarify your own understanding of exactly what some output from ArcGIS really means.

****

GIS paper #1 – "Remote Sensing of Grass Seed Cropping System Components in a Landscape" submitted to *Agronomy Journal* March 2005, returned for revision June 2005 with positive comments, revised and resubmitted Oct. 2005, rejected Feb. 2006.

Revised and submitted to *Remote Sensing of Environment* as "Multi-year GIS of Grass Seed Cropping Practices in Western Oregon" Dec. 2007. Rejected by editor. Revised and resubmitted July 2008. Rejected again by editor.

I had developed a novel approach to get around the 20 band classification limit in ArcGIS and applied it to western Oregon grass seed crops. Editors asked why I didn't just use existing remote sensing classification software like Erdas Imagine. Ultimately I was led to abandon my new approach. This was actually OK, since it turned out that standard software produced slightly more accurate classifications than my method.

I next split the ongoing project (and paper) into two parts, the first one focusing on the ground-truth GIS and the second part focusing on remote sensing classification using Erdas Imagine. Part one was submitted to *International Journal of Geographic Information Science* April 2009 and promptly rejected by the editor before part two could even be sent.

Both the first part and the second part were then revised and submitted to the *International Journal of Remote Sensing* in June 2009, and revised and resubmitted in Oct. 2009.

The remote sensing classification paper was accepted by *IJRS* Dec. 2, 2009, while the multi-year GIS paper was rejected by them Dec. 10, 2009.

GIS paper #1 published – "Remote sensing classification of grass seed cropping practices in western Oregon" by Mueller-Warrant *et al*. 2011. *International Journal of Remote Sensing* 32(9):2451-2480.

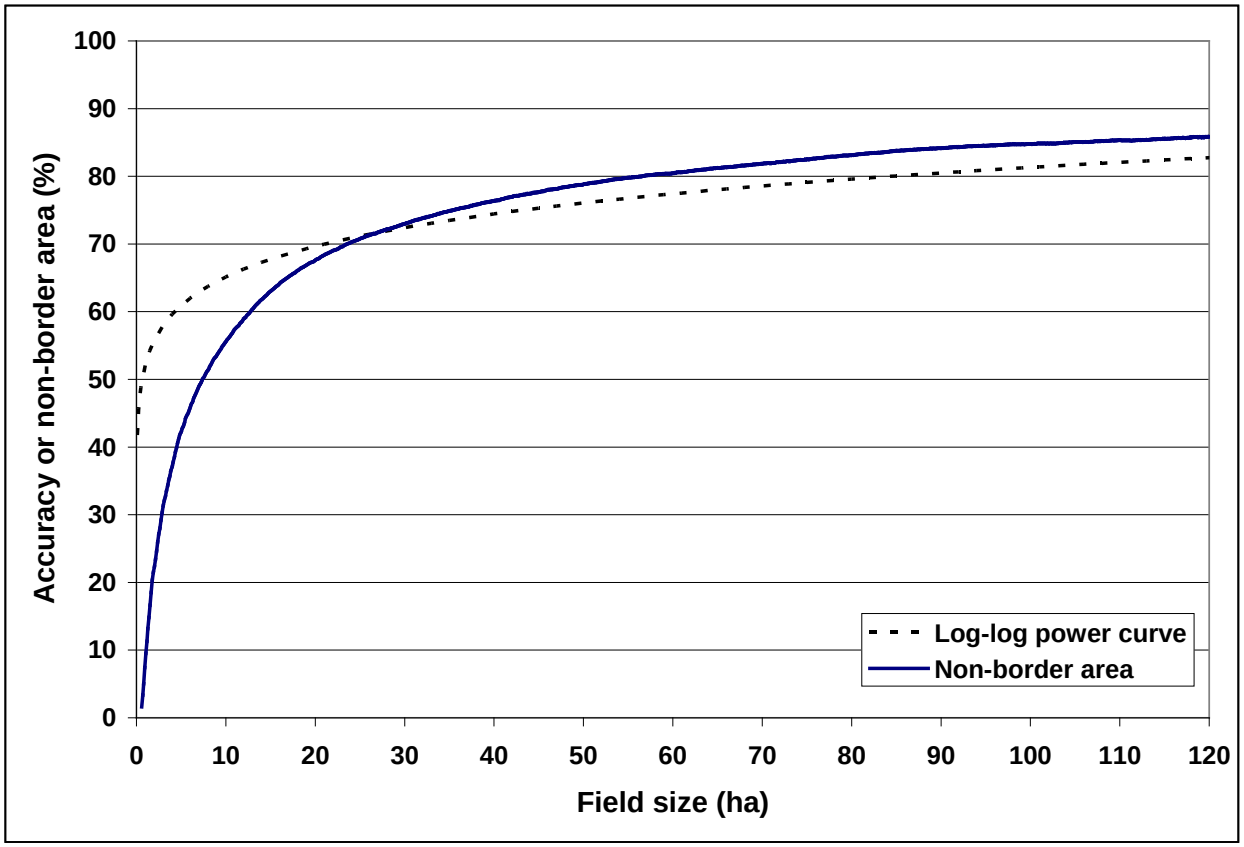
The 3-year version of GIS ground-truth database development and cropping systems analysis has only been published in a series of non-peer reviewed reports to the Oregon seed industry starting in 2002 and as Map ID 736 in the 2003 ESRI-UC. I am now working on a 7-year version of crop rotation history using 56 crops rather than just the 16 crops used in the *IJRS* paper, and may go back to try *Agronomy Journal* again. I may be a glutton for punishment!

The first major statistical challenges was the hard limits within ArcGIS regarding the numbers of classes (20) and image bands (20) that can be used in classification. Many grass seed crops are highly similar, so even all available Landsat images together possessed just barely enough information to separate them.

The second main issue was the gaps in the images caused by clouds and by failure of the Landsat 7 SLC (scan line corrector). We do not have enough 100% cloud-free days in western Oregon to afford the luxury of tossing out partly cloudy images. This issue was resolved by conducting a series of classifications using ever higher number of images within ever smaller areas of (non-gap) overlap. Each pixel was ultimately assigned whatever classification category it received from classifications using the maximum number of available images. Up to 10 steps were used in the 2006-07 harvest year. Number of bands were first increased by calculating NDVI and then reduced by use of principal components analysis from 7 or 8 for Landsat down to 3 orthogonal bands (and down to 2 for MODIS and AWiFS).

The third challenge was how to best convert MODIS 250 m data into 30 m pixels compatible with Landsat images. Part of the *IJRS* publication was a test of whether all 25 of smaller (50 m) pixels within 250 m (MODIS size) should have the same value or whether the corners and edges should be averaged with their (3 by 3) neighbors. Classification accuracy was improved by 3% in training areas and 2% in validation areas when neighboring pixels were averaged. These 50 m pseudo-pixels were then resampled down to 30 m resolution to align with Landsat data.

Another similar issue was how to handle the imperfect alignment between edges of the ground-truth fields and pixel positions/boundaries. One method was to test for identical ground truth values within N by N neighborhoods for N=3, 5, 7, or 9 to remove pixels on the boundaries between dissimilar ground truth polygons. This edging procedure improved classification accuracy up to the N = 7 case, but did so by removing 26, 47, 61, and 72% of the ground truth data (area), including some entire classes by N = 9. Another way to display much the same concept was by graphing the relationship between field size on the x-axis versus either classification accuracy or percentage of the field area that was not on the border.


Accuracy or non-border area (%)
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
110
120
Field size (ha)
Log-log power curve
Non-border area

****

GIS paper #2 – "GIS Analysis of Grass Seed Weeds: I. Crop Rotation and Stand Age Effects" submitted in April 2006 to *Weed Science*. "GIS Analysis of Grass Seed Weeds: II. Patterns of Spatial Clustering and Temporal Change" submitted June 27, 2006, to *Weed Science*.

Both rejected Sept. 11, 2006. Rejection was formally appealed Oct. 6, 2006. Resubmitted Feb. 2007 to *Weed Science*. Revised again and resubmitted Nov. 19, 2007, as a single paper with some topics deleted. Revised one final time Mar. 4, 2008. Accepted for publication Apr. 21, 2008 by *Weed Science*.

GIS paper #2 published – "GIS Analysis of Spatial Clustering and Temporal Change in Weeds of Grass Seed Crops" by Mueller-Warrant, Whittaker, and Young. 2008. *Weed Science* 56:647-669.

Challenges prior to journal submission

Granted access to non-spatial database of seed certification records (going back to 1994) in early 2002. Downloaded by county and year, then developed Excel spreadsheets, an ArcGIS personal geodatabase, RELATE links between tables to draw (edit) tentative field locations based on Township/Range/Section position, field acreage, road maps, Landsat NDVI, and written directions to the fields given to the inspectors. I limited the georeferencing project to Linn County, the "heart of Oregon's grass seed industry." I also promised growers and OSU Seed Certification to keep field locations and all inspection data confidential. Displayed first maps in Dec. 2002 at the annual OSGL conference in Portland, using Inverse Distance Weighting to "spread" weed severity scores or simple presence/absence data across the entire Willamette Valley, distributing the grass seed fields as points randomly positioned within the actual TRS (mile square) general locations. By the time I finished fully "spatially enabling" the Linn County field inspection reports, I had data on 10,643 harvests from 2,779 fields over a 10-year period (1994-2003). Up to 43 different weed species were either absent from individual fields or rated as being present at *trace, many,* or *excessive* levels. I initially coded those ratings with values 0, 1, 2, or 3, but eventually converted them to 0, 0.75, 1.75, 2.75, and 3.75 (when species was grown as a crop) to more closely approximate "normally distributed" data.

Letter from OSCS granting permission to publish:

June 14, 2006

Hello, George

We agree that your information reasonably protects grower identity. We appreciate your efforts and consideration in this regard.

Thanks,
Sandy

Oregon Seed Certification Service

Letter to Weed Science formally appealing rejection:

Dear Robert E. Blackshaw:

We hereby formally request reconsideration of your recent decision to reject manuscripts WS-06-035 and WS-06-111. In the following paragraphs we will conclusively demonstrate that Reviewer # 1 was in error in most of his or her criticisms of both papers. We will do so by following each critical comment made by the reviewer (*in italics*) with our rebuttal (**in bold type**). If you agree to reconsider either or both of these manuscripts, we will be happy to work with your editors regarding all technical points raised in the reviews. But given the condescending tone of the criticisms by reviewer #1, we see little point in attempting to revise and resubmit without admission on your part that the major criticisms by reviewer #1 were invalid. Since the associate editor had deferred to the recommendation of reviewer #1 over that of reviewer #2, we presume that favorable action on your part to our request for reconsideration would of necessity include a change in the associate editor's recommendation. If that indeed happens, we will be happy to respond in detail to the specific points raised by the associate editor. The following rebuttal only deals with comments made by reviewer #1.

Sincerely,

George Mueller-Warrant on behalf of himself and the co-authors

In the following pages, the reviewer's comments are italicized and are followed by my arguments against their position in regular type.

<u>WS-06-035:</u>

Contrary to the authors' opinion, the GIS analysis is not needed for any of the objectives. All that is needed is knowledge that data came from the same field for different years.

As we stated on lines 109-111, GIS methods were indeed required to achieve specific objectives 1, 2, and 4. The reviewer's cavalier statement that "*all that is needed is knowledge that data came from the same field for different years*" indicates his or her unfamiliarity with details of the record keeping procedures used by the Oregon Seed Certification Service. The OSCS field identification scheme has no built-in method to carry over identities of fields from one certified planting to the next. Instead, the OSCS relies on growers to self-report the last five years of field history in their applications for seedling inspection of new stands.

Even this information, however, was not available on-line when our GIS was constructed, only existing then as tens of thousands of paper forms stored in filing cabinets. … If the grower-reported field history

had been available online to us, he or she would be partially correct, subject to the untested accuracy and honesty with which growers filled out their field history. But as it is, he or she is quite mistaken. Objectives 1, 2, and 4 were impossible to achieve without the GIS. It would certainly be to the OSCS's advantage to manage the certification and inspection of fields using a GIS, but that is not the current situation.

The information on weed severity could be of interest to readers of Weed Science if the data could be analyzed statistically. The author's regression analysis is not appropriate because weed severity is an ordinal variable.

We did indeed conduct statistical tests viewing weed severity data as simply categorical rather than ordinal or interval data in chi-square tests of independence of weed carryover tendency in Table 5. Did reviewer #1 miss seeing the first two columns of that table? Observed probabilities for rejection of the assumption of independence between stands (chi-square test of categorical data) were very similar to those for correlation and linear regression from the analyses of severity between stands as interval data.

The reviewer apparently rejects our conversion of weed severity scores made by the field inspectors (none, trace, many, excessive) to logarithmic interval data (0, 1, 2, 3). In so doing, the reviewer is also implicitly rejecting most of the visual ratings of weed control, crop injury, lodging, and plant disease severity published over the past 50 to 100 years in the fields of agronomy, weed science, and plant pathology, whether or not they were normally distributed (or could be transformed into normally distributed interval data). All that is needed for validity of our conversion of weed severity from categorical to interval data is satisfaction of the equal spacing requirement that similar differences between pairs of observations have similar meaning for all pairs.

…

Chi-square statistical tests of independence in Table 5 viewing the data as simple binary presence/absence values produced essentially identical interpretations of the tendencies of individual weeds to carry over from one stand to the next as did the correlation and regression analyses of the interval data. Reviewer #1 is welcome to question whether or not our conversion of weed severity scores to interval values was done correctly, but the similarity of our results between categorical analysis and correlation (or regression) combined with our knowledge of the approximately exponential differences in absolute weed density (and seed production) between categories suggests the reviewer has certainly failed to prove his or her claim the data could not be "*analyzed statistically.*"

Data is presented and compared without statistical analysis (for example, Table 6) or at least, an indication of the variability. If there are not enough cases, for a statistical analysis, then these should be deleted.

The data underlying the means presented in Tables 6 and 8 failed chi-square tests of independence of weed species and planting date (Table 6) or stand age in crop years (Table 8), demonstrating that some statistically significant interaction occurred between the classification factors structuring each table. We regret failing to footnote both tables with this overall observed probability (P<0.0001). Inclusion of standard errors in Tables 6 and 8 is certainly one approach we could take to answering reviewer #1 criticism, but doing that would risk making those tables even larger and harder to read.

Reviewer #1 Evaluations:
Recommendation: Manuscript unsuitable for publication in present form. Reorganize, rewrite, and resubmit.
Originality: Fair
Technical quality: Poor
Clarity: Poor
Importance: Fair

We believe that Reviewer #1's evaluations of this manuscript's Originality and Importance as only *Fair* are based on his or her misunderstanding of the capabilities and limitations of the non-GIS OSCS database. Three of our five objectives could not have been achieved without creation of the GIS that allowed us to match up data from different plantings conducted over time within the same field. We suspect that many similar sources of potentially geo-referenceable data exist within the files of regulatory agencies, private industry, and weed science researchers throughout the world. … We believe that Reviewer #1's evaluation of this manuscript's Technical quality as *Poor* was based on his or her failure to recognize our analysis of the key data in Table 5 viewed from the perspective of categorical variables as well as interval variables to which he or she objected. We also believe that our conversion of the data to an interval variable is fully justified and statistically rigorous. Weed Science's decision to reject this manuscript has aborted the ability of the manuscript review/revision process to improve any technical weaknesses in the manuscript and enhance the understanding with which the reviewer will approach future manuscripts on similar subjects.

WS-06-111.

For several reasons, I could not evaluate the scientific value or significance of the results and consequently choose between rejecting or rewriting and re-submitting for review:

1. I could not follow authors' manipulations of spatial locations of fields for "privacy reasons." Consequently, I could not evaluate whether these processes would affect the integrity of the data for spatial analysis. I suspect it does. Can privacy be maintained by shifting geographic coordinates (such as moving the fields to China)? Another difficulty with the requirement for privacy is that a map of field locations is necessary because clustering of sample locations (field locations) affects conclusions of most spatial analyses.

Reviewing #1 is mistaken in his or her speculation that "*manipulations of spatial locations of fields for privacy reasons"* **affected the "***the integrity of the data for spatial analysis***." All statistical and spatial analyses were conducted using the most accurate data we possessed regarding field locations, borders, and centroids. The only thing that was degraded in resolution to protect grower privacy was the public map of Gi* probability. The Moran's I and Getis-Ord G spatial statistics are designed to inherently handle the issue of possible clustering of sample locations. Evidence that they succeeded in doing so can be seen within our own manuscript in that while most weeds clustered significantly, not all of them did.**

2. I am not convinced of the necessity of the authors' use of an "offset" of field locations and again, could not follow the author's procedures for the "offset" well enough to evaluate how these processes would affect the integrity of the data for spatial analysis. The authors show that choice of the offset affects the results of Moran's autocorrelation (not a worthwhile result for a paper - what is interesting about Moran's I is how the value changes with choice of lag distance). I think arbitrary choice of an offset is as "capricious" as "excluding pairs of points closer than some designated distance," especially since Moran's I and hot spot analysis depend on relative location. There are alternatives. Maybe centroid is not the best method to designate location of field (centroids may not always be located in a polygon)? Are some of these actually the same fields and should be designated as such

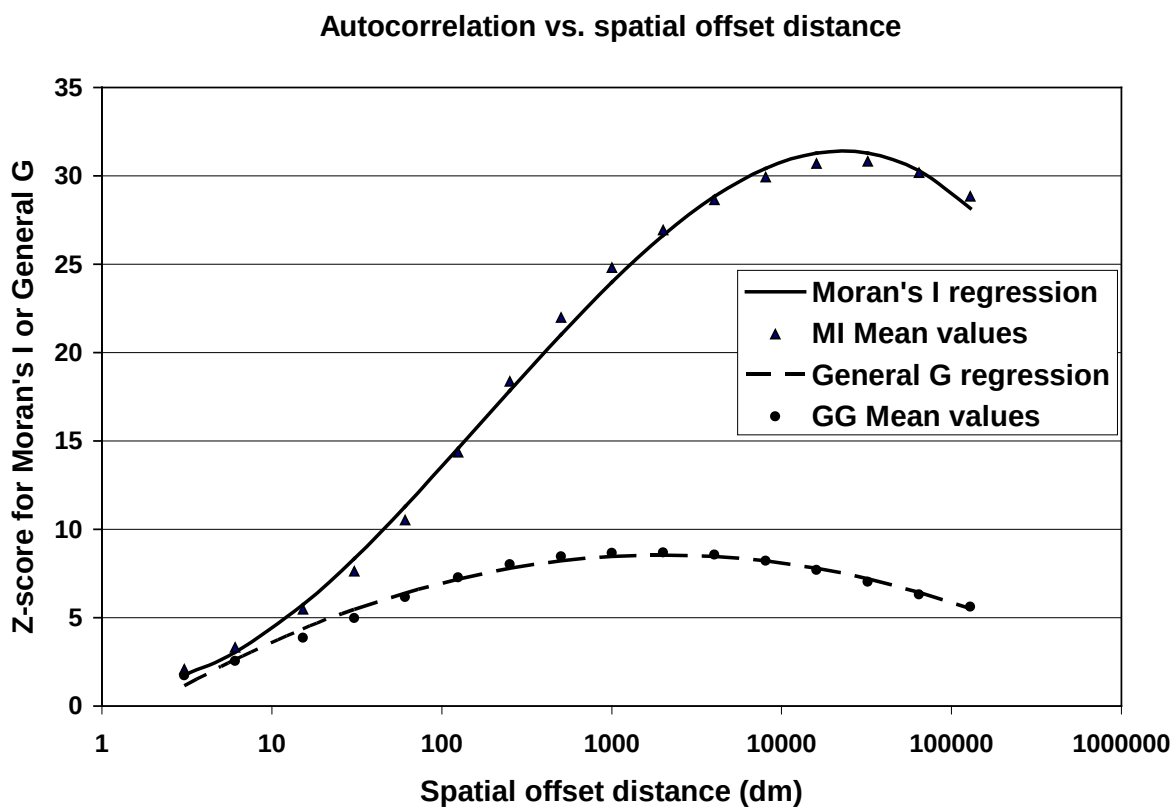
How are the offsets influencing the choice of lag distances for the Moran's I analysis? If the offset is a reasonable approach to what the authors' consider a problem, there should be a defensible method for choosing an appropriate offset rather than presenting all those results!

In Moran's I spatial autocorrelation (and Getis-Ord General G high/low clustering), numerous options exist for what reviewer #1 refers to as "*choice of lag distances*", or to use ESRI's terminology, distance weighting matrices. One option is to pick an arbitrary distance within which two points are considered to be neighbors and beyond which they are not. Table 5 provides an example of results from using this 'fixed distance band' distance-weighting technique. To determine distance for peak spatial autocorrelation (or clustering) or maximum range of any significant spatial autocorrelation, it was necessary to run Moran's I or Getis-Ord General G repeatedly at increasing lag distances, ranging in our case from 1 km out to 59 km. Table 5 only presents results for the two most interesting distances (peak and maximum range of significance), but dozens of analyses were run in batch file mode (Python

scripting) for each weed species at increments of 1 km near both the peak and the maximum range, and at wider increments while searching for approximate locations of the peak and the maximum range. There was no need to use our 'spatial offset' in this distance weighting method, as everything within a given distance had a weight of 1 and everything beyond it had a weight of 0.

The 'spatial offset' came into use when alternative distance weighting methods such as inverse distance (the ESRI default) or inverse distance squared were used to 'conceptualize the spatial relationship.' Without addition of a 'spatial offset', these methods were highly sensitive to the presence of a few unusually close field centers, cases where field shapes had changed over the 10-year period of our data and the fields partially overlapped. …

[Figure (below) added to revised paper summarized data that had been present in two separate tables. The points on the graphs are averages for 43 weeds showing mean changes in Z-scores of Moran's I and General G as spatial offsets were increased in an IDW model.]



So the question became how to fix this artifact in a defendable, repeatable method. One cumbersome method would have been to go back to the field polygons, find the ones whose centroids were closest, and move those centroids so that none of them were ever any closer than such critical value, such as the average field diameter (or N-S extent, or E-W extent), or some fraction thereof. This would have solved the problem of getting drastically different Moran's I results for 'inverse distance' (Tables 3 and 4) versus 'fixed distance band' (Table 5) methods, but could hardly be called ideal. What we did instead

was to force the distance between all pairs of fields (when using inverse distance weighting) to differ by an arbitrary additional amount we called the 'spatial offset'. We varied this 'spatial offset' over a wide range of values, only half of which were presented in Tables 1 and 2), to identify the adjustment factor that gave us the clearest view of spatial autocorrelation in our data. Rather than being "*as capricious" as 'excluding pairs of points closer than some designated distance,'"* our method is defensible, repeatable, and easily extended to other settings. It also has the advantage of avoiding exclusion of any data. We found that a 'spatial offset' of 402 m worked well for both Moran's I and Getis-Ord General G, although distances up to four times greater or smaller than that were almost as good. A distance of 402 m would correspond to a field size of 40 acres, which is not that different from the average size of certified grass seed fields. What we have proposed is a general solution to a problem that many other researchers will likely encounter in future attempts to characterize spatial clustering of their own data.

Other concerns if the authors choose to resubmit the paper:

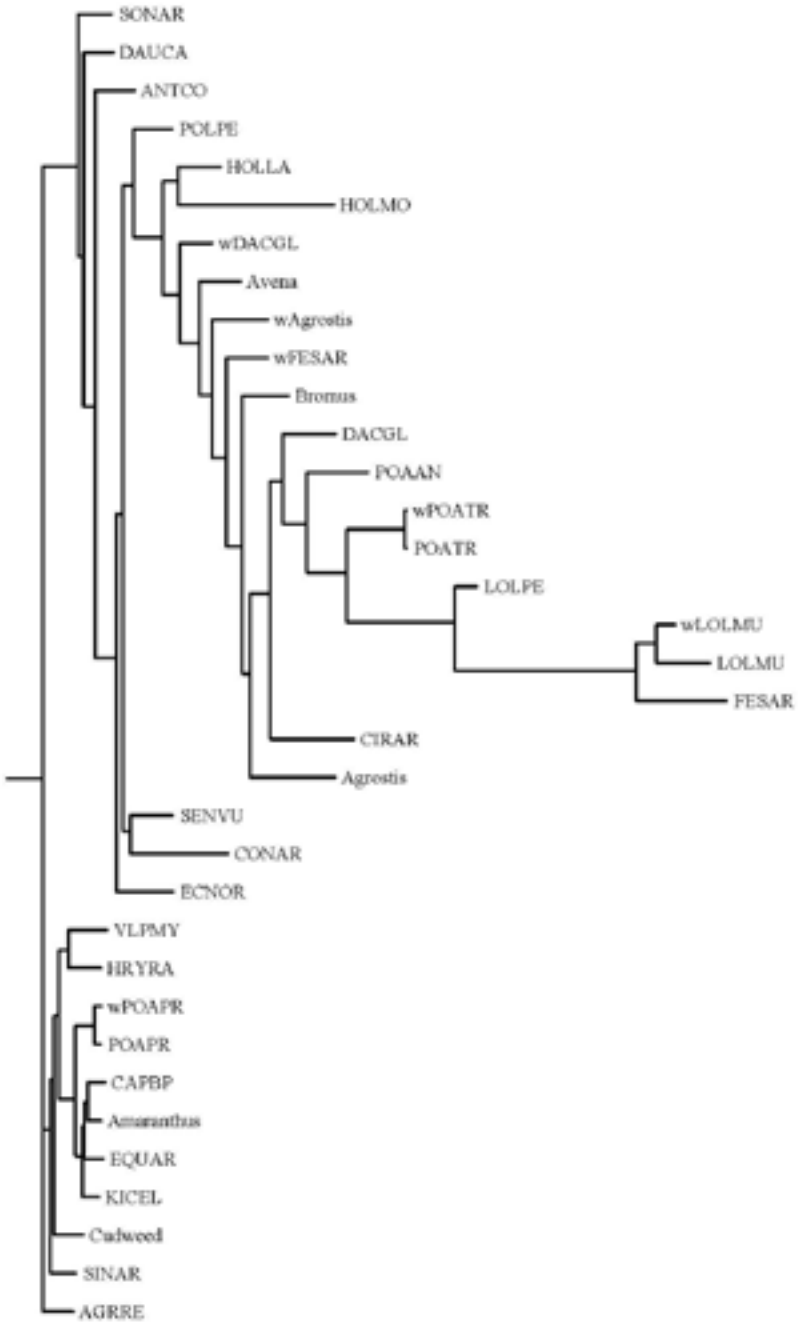
In general, the results of the spatial analyses describing spatial continuity with relation to distance may not be meaningful, or at least, very difficult to interpret. Consider the difference of interpretation between sampling weed density within a field and sampling weed density among fields. The scientist studying spatial continuity of weed density within a field chooses (somewhat) random sample locations and expects analysis based on those sample locations to be similar to results from another set of sampling locations. The set of sample locations (fields) are constrained by houses, roads, rivers, growers' preference, etc., not just the scientist's choice. Actual spatial location is critical in the interpretation of these results.

Reviewer #1 raises some interesting points ... Going on to the subject of "*actual spatial location*" of fields being "*constrained by houses, roads, rivers, growers' preference, etc."* the first point we need to emphasize is that we are not dealing with just a sample of the underlying population, but instead have the entire population of Linn County certified grass seed fields. So, do the distances between field centroids (with or without addition of our 'spatial offsets') show artifacts or biases introduced by the constraints mentioned by reviewer #1? The short answer is no, with the major reason being the extremely large number of fields in our dataset and the relatively large distances over which spatial autocorrelation occurred. If we had only used a few fields in our analyses, then certain distances (e.g., integer multiples of a mile due to typical spacing of roads) might have been unusually common and potentially able to distort the spatial statistics. But with the average distance for peak Moran's I spatial autocorrelation being 13 km, the average field size being 18 ha, the average distance between fields being 92 km, and the total population being 2779 fields, small numbers of neighboring fields around an individual field (the situation in which the constraints on location are most likely to cause problems) were extremely rare events. On lines 433-436 we discussed the subject of how many neighboring fields affected the Gi* probability calculations, and the average number across all weeds was over 600. In that paragraph we also discussed the steps we took to insure that even on the edge of the data we knew that we had at least seven neighbors around any point for any weed species in Gi* calculations (and a minimum of 12 neighbors when converting Gi* point location values into rasters).

I did not review the section on the soil effect because this part of the manuscript should actually be a separate manuscript. That manuscript should include a discussion of why the Fitch-Margoliash tree diagrams and principal component analysis are meaningful in this context and an appropriate literature review. Also, important results should be described in the abstract.

If reviewer #1 had reviewed the soil effect section (lines 364-431; Tables 8, 9 and 10; and Figures 1 and 2), he or she would have found that the figures provide some of the "*better*" ways to "*summarize*" our "*data*" and "*illustrate consistent patterns*" that he or she complained were lacking. Figure 1 in particular provides a rich summary of the similarity (or differences) in spatial distribution patterns of the 35 most prominent weed species in our database. We are disappointed that reviewer #1 failed to consider this figure – it does more than anything else in our paper to extract a key central feature of the gigabytes of data in our GIS and compress it into a single image the human mind can comprehend.

[Figure (below) is a Fitch-Margoliash dendogram showing the distance between 35 weed species in terms of 36 soil types over a 10-year period.]


SONAR
DAUCA
ANTCO
POLPE
HOLLA
HOLMO
wDACGL
Avena
wAgrostis
wFESAR
Bromus
DACGL
POAAN
wPOATR
POATR
LOLPE
wLOLMU
LOLMU
FESAR
CIRAR
Agrostis
SENVU
CONAR
ECNOR
VLPMY
HRYRA
wPOAPR
POAPR
CAPBP
Amaranthus
EQUAR
KICEL
Cudweed
SINAR
AGRRE

Reviewer #1 Evaluations:
Recommendation: Manuscript unsuitable for publication in present form. Reorganize, rewrite, and resubmit.
Originality: Good
Technical quality: Poor
Clarity: Poor
Importance: Fair

We believe that Reviewer #1's evaluation of this manuscript's Technical quality as *Poor* are based on his or her failure to adequately grasp the central data quality issue we encountered in our initial analyses of spatial autocorrelation of the maximum weed severity observed over the full 10-year period of our data, i.e., the domination of the default inverse distance weighting matrix by a very small number of pairs of fields whose centroids were much closer to each other than was reasonable, indeed often falling inside each other's polygon. We devoted two whole tables to description of methods we devised to deal with this problem in a general, consistent, and non-arbitrary way. We did so based on our firm conviction that other users of GIS software who attempt to characterize and summarize patterns of spatial autocorrelation will encounter similar problems in their own datasets. In regards to reviewer #1's evaluations of this manuscript's Clarity as "*Poor*", we agree that its clarity could be improved. We suspect, however, that much of reviewer #1's problem with our manuscript's clarity was caused by his or her focus on finding perceived problems with our technical methods. If the reviewer had initially perused it in a 'non-combative' mode starting with the figures and then working back to the tables and methods he or she would likely have had fewer problems with our manuscript's clarity. His or her own admission that he or she did not even bother to review the soil effect section is troubling, as Figure 1 (associated with that section) is a single visual summary of the relationships among 35 weed species over 10 years and 2779 individual fields. Reviewer #1 simultaneously bemoans a relative shortage of simple summaries of what our data mean while arguing with details of our methods. We find it hard to imagine that he or she would really have been any happier if our methods had been compressed to a single paragraph and our summaries expanded. We do think that the readability of the manuscript could be improved moving some of the materials & methods and some of the tables to an online supporting status where it wouldn't distract the casual reader from just finding the results, if Weed Science employed that system. Unfortunately, our attempts to do just that with the Python source code for running the geoprocessing tools (noted on line 143 as being available online at http://www.ars.usda.gov/pandp/people/people.htm?personid=4006) failed to provide reviewer #1 with the details he or she claimed to need to evaluate the validity of our geoprocessing methods (failing because he or she apparently didn't bother to access the link).

Apr. 21, 2008 letter acceptance for publication by *Weed Science*:

Dear Dr. Mueller-Warrant,

I am pleased to inform you that your revised manuscript "GIS Analysis of Spatial Clustering and Temporal Change in Weeds of Grass Seed Crops" has been accepted for publication in Weed Science. Your manuscript will appear in the next available issue. The month of publication will appear at the bottom of the page proofs, which you should be receiving from the printer in about 10-12 weeks.

Thank you for your contribution to our Journal. If you have any questions, feel free to contact Tracy Candelaria, Managing Editor, by e-mail at wssa@allentrack.net.

Dear Dr. Mueller-Warrant,

I thought you might like to know that as I read this revision I had many "Aha" moments where I said, "That's what this means!". You have done an outstanding job of making this paper readable and understandable to the numerically literate. In addition, the interpretations you have put to the statistics make this paper a valuable contribution to GIS and weeds research. Numerical literacy will not be required to glean considerable useful information from this paper.

Best Regards,
Paul Watson

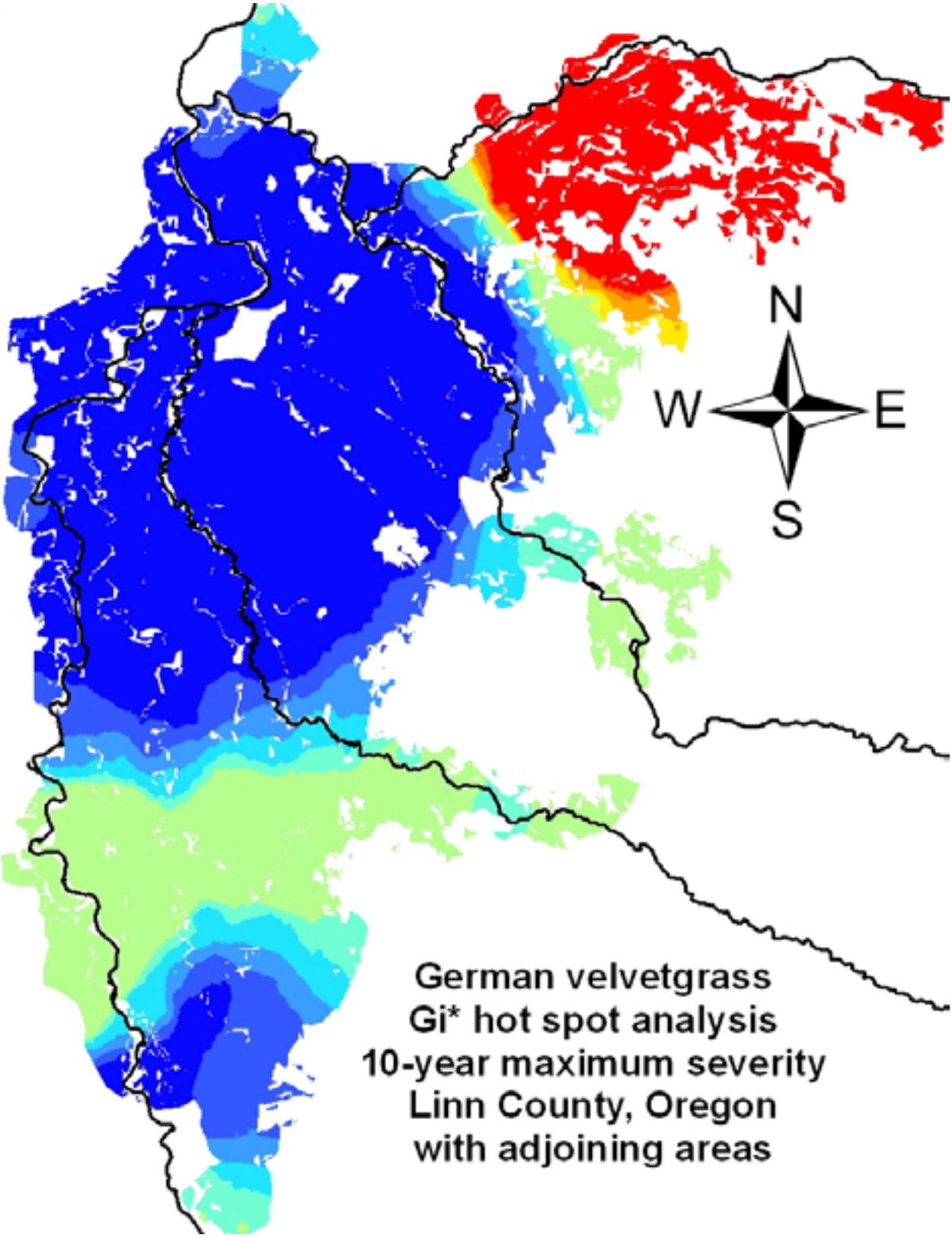
I know that this has been a long process. Thank you for your patience and work on this paper.

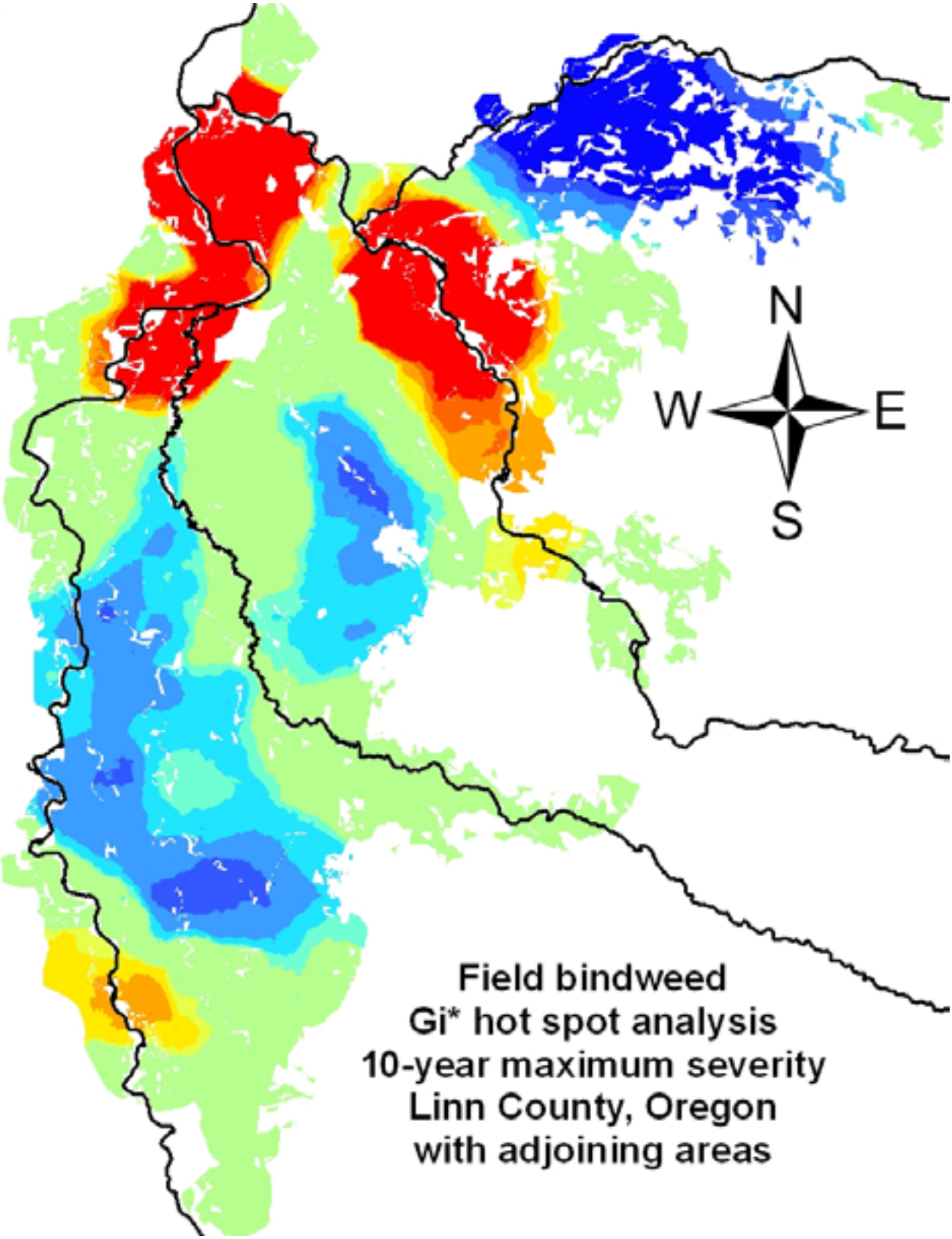
Sincerely,

Robert E. Blackshaw
Editor, Weed Science

Amongst the things we got to add to paper were equations for global and local versions of Moran's I autocorrelation and General G high/low clustering, along with descriptions of their statistical properties.

We also got to add PCA, RDA and CCA multivariate statistics to our paper, generating two more tables and four more figures.




N
W
E
S
Field bindweed
Gi* hot spot analysis
10-year maximum severity
Linn County, Oregon
with adjoining areas

****

GIS paper #3 – "Geospatial identification of optimal straw-to-energy conversion sites in the Pacific Northwest"

First step was sending in abstracts (and then complete versions) of two proposed papers to the editor of *Biofuels, Bioproducts, and Biorefineries* to ask if the journal would be interested in a formal submission of either or both of the papers. The editor replied that she would be interested in the second one but not the first.

GIS paper #3a – "Spatial distribution of cereal and grass seed straw as potential bioenergy feedstock in the Pacific Northwest" by Mueller-Warrant, Griffith, and Whittaker. Submitted Nov. 4, 2009, to *Remote Sensing of Environment*. Rejected by editor, revised and resubmitted Nov. 27, 2009. Rejected again on Dec. 6. Revised and submitted to *International Journal of Remote Sensing* Feb. 2, 2010. Rejected May 10, 2010.

GIS paper #3b – "Geospatial identification of optimal straw-to-energy conversion sites in the Pacific Northwest" by Mueller-Warrant, Banowetz, and Whittaker. Submitted Nov. 5, 2009, to *Biofuels, Bioproducts, and Biorefineries.* Revisions requested Feb. 1, 2010. Revised manuscript submitted Mar. 19 and paper accepted Apr. 1, 2010.

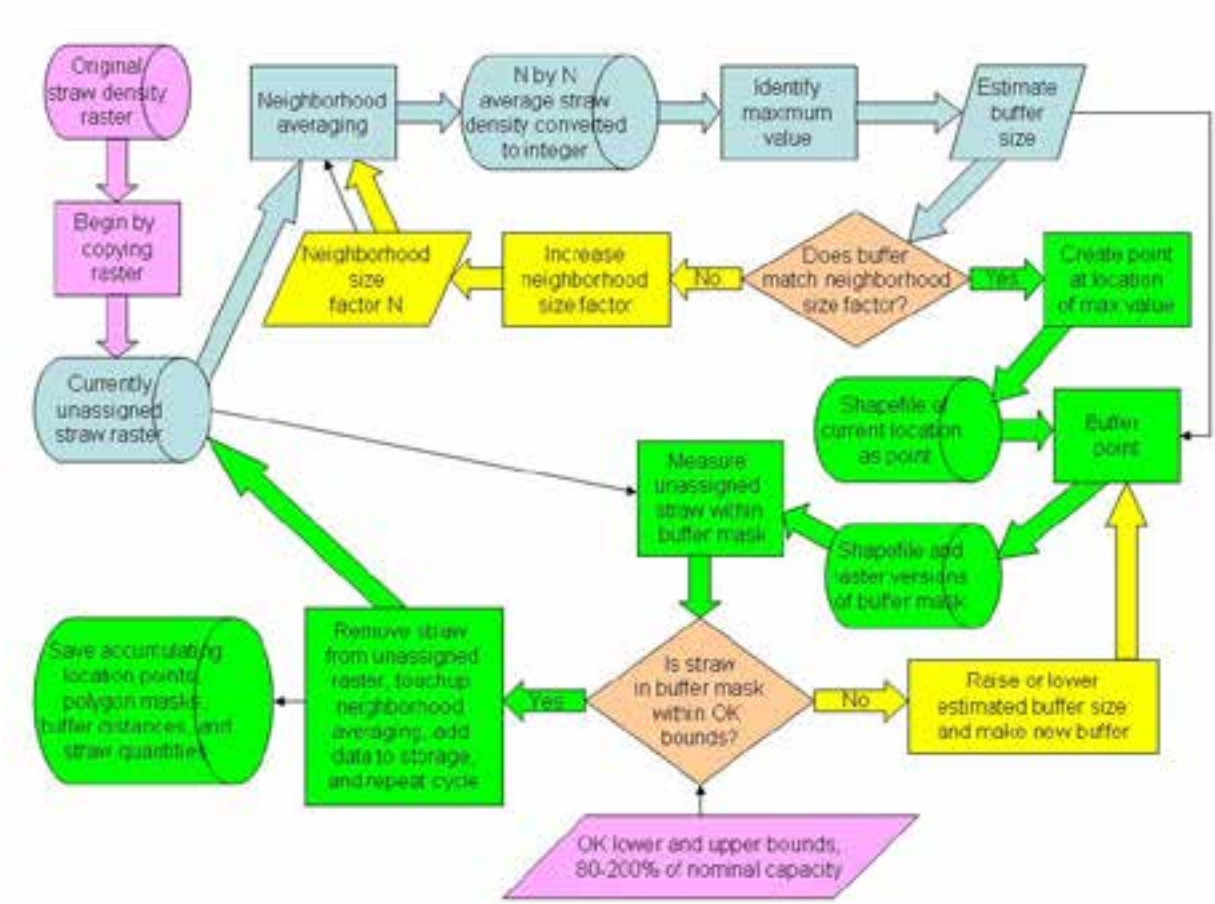
GIS paper #3 published – "Geospatial identification of optimal straw-to-energy conversion sites in the Pacific Northwest" by Mueller-Warrant, Banowetz, and Whittaker. 2010. *Biofuels, Bioproducts, and Biorefineries* 4:385-407.

Two major issues were tackled in this project. The first was to convert NASS CDLs from three different years (plus the western OR GIS) – ID in 2005, WA in 2006, and OR, WA, and ID in 2007 – into yearly data on cereal crop locations across the entire PNW. Combined with per acre (by county) yield estimates, harvest indexes, and NRCS ground cover/straw residue requirements we generated 250 m resolution rasters for available straw over multiple years in the PNW. Unfortunately no one has wanted to publish those details. One editor implied that he might be interested if I wanted to expand it to cover many more years (and maybe many more states).

The second major issue was the central point of the manuscript that did get published – how to define optimal locations for biofuel plants of varying capacities. We proposed that entrepreneurs would be more likely to follow a 'lowest-hanging-fruit' approach than global optimizations such as K-means that are really designed for a centralized control model where the government or a monopolistic business needs to figure out how to provide all the customers with a service such as medical ambulances or cell phone coverage. SSAO (Series of Sequentially Approximated Optimizations) started by assuming that some particular neighborhood size (N) might contain enough straw to meet to biofuel plant's capacity. The entire PNW raster was then processed using NxN neighborhood averaging, the maximum value found, the buffer size estimated that would supply enough straw, and that size compared with the neighborhood. Once N was large enough to supply the straw, points and polygons were created to measure the actual total quantity of straw within a tested range. If too much or too straw was found,

the buffer size was lowered or raised and tested again and again until straw was between 80 and 200% of plant capacity. That straw was then removed from the total available raster and the whole process repeated until there was no straw left or the neighborhood approached the entire PNW. Approximately 6,200 of the smallest-sized plants (1 million kg/yr) were needed to process all the straw. Not only did this mean 6,200 trips through the loop, but the testing we conducted also ran it for 3 different individual years plus 2 averages. Some of the runs took ~1 week to finish on a single computer.

Procedural steps in our method:



Several of the questions we were asked during review of this paper concerned what were the statistical properties of our method and of its results. One of the more interesting findings we added came from comparing the order of siting of the plants and the size of the required buffers to supply the straw. Spearman's rho for these rankings was 0.994, 0.942, and 0.978 for the 1, 10, and 100 million kg straw per year plants. I will leave the rest of the statistical properties of SSAO to the paper itself, and answer any questions if there is any time remaining. But not before I show one more figure from the biofuel paper.