Correspondence with Dr. Brian Kinlan, NOAA National Ocean Service Contractor, CSS Inc. NCCOS-CCMA-Biogeography Branch, regarding GIS hotspot analysis using survey data

Hi Andy,

Thanks for your note, glad you found our advice helpful. I think the hurdle model approach, with transformation of non-zero data, will probably be useful for you. To implement this in ArcGIS, you may want to look into Geostatistical Analyst (for handling spatial autocorrelation) and MGET (http://code.env.duke.edu/projects/mget), which brings R functionality into ArcGIS.

It sounds like you are on a pretty tight timeline, and we are flat out on existing projects here for the next few months, so we can't commit to an immediate collaboration. Of course, we are always happy to help with an occasional email or phone call--and hopefully Charlie and I have given you some useful starting points for the next few months of your work.

That said, thinking about the medium to long term, we are definitely interested in this type of work. Motivated by management information needs, a core focus of our group's research has been on developing robust, peer-reviewed statistical methods for predictive hotspot modeling and accuracy assessments. We would be open to collaboration down the line if these types of questions continue to be important to your group. If you think that is a possibility, we could discuss how best to build that collaboration. I think my boss Chris Caldow (cc:ed here) would be happy to hop on the horn with you at some point to discuss how that could work.

Sorry we can't be of more direct help at this point, but hopefully now that this connection is made we'll have an opportunity to come together at some time in the future and work in a more formal way together on these types of analyses.

cheers,
Brian
On 1/22/2013 2:48 PM, Andrew J. Applegate wrote:
Brian,

Thanks for the expert advice. I am the chair of the Council's Closed Area Technical Team (CATT), which Laurel and Sean are members. My colleague Tom Nies and I have been working with the Moran's I and Ords-G* statistical procedures to identify hotspots of spawning (i.e. large) and juvenile fish for the CATT. Some of the special statistical challenges that we have encountered led to the email thread below.

While we are capable of doing a power transformation using the Box-Cox procedure, none of us here are really adept at developing a ZINB model in R. We were initially concerned that omitting no catch tows might bias the hotspot analysis, some of those concerns have been addressed through sensitivity and simulation analysis. Others however are a little uncomfortable leaving zero tows out of such an analysis, however. We also have the data structured in such a way that tows not catching a species (cod, for example) can be treated differently from tows not

catching large cod (50+ cm cod, for example). These latter type of tows might be treated differently in a hurdle model approach, which coincidentally was where we had started out with this analysis.

I think we also need to keep in mind what we are trying to achieve and the minimum amount of statistical rigor (and hence time) that is necessary to achieve our objective. It sounds, based on your advice, that the hurdle model approach would be sufficient and that transforming the positive catches via the Box-Cox procedure would be the way to go. I also have to keep in mind that the more complicated a model is, the more difficult it is to sell to managers (Council members) and the public.

I appreciate your advice and would be willing to collaborate, doing most of the work ourselves and turning to you for guidance. To do this, I could provide you with an example GIS feature class (data layer) and then we could talk by phone or collaborate via GoToMeeting, after reading the NY tech memo.

Does this sound like something that you could do with us?

PS. I am using ArcView 10.0 and 10.1 (with all extensions), and I have experience using Systat for statistical analysis. I think Michelle (a staff member) also has some experience and access to R, although she probably has no or little experience with ZINB models.

Andy Applegate
NEFMC staff
978-462-0492 ext 114

From: Brian Kinlan (NOAA Affiliate) [mailto:brian.kinlan@noaa.gov]
Sent: Tuesday, January 22, 2013 11:53 AM
To: Sean Lucey
Cc: Charles Menza; Randy Clark; Laurel Smith; Andrew J. Applegate
Subject: Re:

Hi Sean, Laurel, et al,

Yes, I think we may have met at the NART Seascape Ecology workshop up at Rutgers last year? Thanks for dropping a line. I actually think this question may have already reached us via a request from Laurel Smith to the NMFS-GIS list. Below I've copied the email thread. My response is in blue and my colleague Charlie Menza's response in red. The bottom line is we have worked a lot with this issue and have some good solutions, but it is not a simple problem.

Your best bet, if you don't have a hard-core spatial statistician on board and/or don't have the money to engage one through collaboration, is probably to bin the data at a spatial resolution that is coarse enough to eliminate most spatial autocorrelation, and then model presence probability using indicator transformed (0/1) data and CPUE conditional on presence using only non-zero tows, and a box-cox transform to acheive approximate normality. This approach is called a

"hurdle model", "delta-transgaussian", or "two-stage" model described in our NY seabird work cited below.

However, this hurdle model approach involves an approximation of CPUE as a continuous quantity (box-cox transformed to be gaussian) when it might be better treated as discrete using a count distribution. Another option, then, is to use count distributions such as a zero-inflated negative binomial (ZINB) regression, which you can implement in R. We are currently taking the ZINB approach for seabirds in the mid-Atlantic, using boosted regression trees which can handle non-linear functional responses and high-order interactions, and having good success with those. You could also apply the same ZI approach with a box-cox transformed gaussian instead of the NB, treating CPUE as continuous. This would improve on the hurdle model by distinguishing between "structural zeros" (true absence of the species) and "statistical zeros" (places where the species wasn't observed simply as a result of random variability inherent in the count process). The hurdle model described above assumes that all zero tows are structural (the species is truly absent), which can be problematic if your temporal window is small (but if you have a long-term dataset and are only interested in the climatological average, as we were in the case of the NY seabird work, it may be an acceptable assumption) or if low detectability/catchability is an isssue.
I hope this helps a bit. You will find many helpful citations re: the zero inflation and two-stage model approaches in our NY tech memo cited below. Feel free to contact us if you'd like to discuss possibilities for collaboration on this.

cheers,
Brian

--
****
Brian P. Kinlan, Ph.D
Marine Spatial Ecologist

NOAA National Ocean Service
Contractor, CSS Inc.
NCCOS-CCMA-Biogeography Branch
1305 East-West Hwy, SSMC-4, N/SCI-1, #9224
Silver Spring, MD 20910-3281

301.713.3028 x-157 (Tel)
301.713.4384 (Fax)
http://ccma.nos.noaa.gov/about/biogeography/
****

-------- Original Message --------
Subject: Re: Fwd: geo-spatial statisitcs
Date: Fri, 18 Jan 2013 15:27:15 -0500
From: Randy Clark - NOAA Federal <randy.clark@noaa.gov>
To: Charles Menza - NOAA Federal <charles.menza@noaa.gov>
CC: Brian Kinlan (NOAA Affiliate) <brian.kinlan@noaa.gov>, Chris Caldow - NOAA Federal <chris.caldow@noaa.gov>

thanks guys. I passed on Brians message and offered myself to chat more general things if they want. I definitely let them know we have dealt with this issue several times and know you can get over your head quickly so I advised them to seriously think about what they want to do and call us if they need help. We could talk potential collaboration at that point. Thanks for your time to answer and I'll let you know if anything happens out of this.
rc

On Fri, Jan 18, 2013 at 3:18 PM, Charles Menza - NOAA Federal <charles.menza@noaa.gov> wrote:

Randy,

We struggle with developing unbiased maps with the same effort issues. And as Brian mentioned there is no simple solution. The solutions he has identified are best. In our work most of the statistical work is done in Matlab or R. We have done some kriging in ArcGIS, but found it very cumbersome with large datasets. I do no have experience with MGET, but it might be useful as well.

I have some time to help, but I'll only be useful for GIS and basic statistical information. For most modeling and complex statistical answers they would need someone with more modeling experience than me.

One other solution to consider, and which is similar to one mentioned by Brian, is to bin the samples into a continuous network of grid cells and show sample statistics for each grid in a map. A decision is needed to determine how much information constitutes sufficient information to show reliable data. For instance, it probably isn't acceptable to compare a bin with 1 sample against another bin with 200 samples. The grid dimensions can be adjusted to balance data retention and usefulness. A mask can be overlaid on the output to show certainty based on how many samples are in each grid. The drawbacks of this method are that the output is coarser, information is lost in cells with too few samples, and sampling bias still remains an issue (although sampling bias is generally an issue when translating a stratified design into another analytic spatial framework).

Good luck,

Charlie

On Wed, Jan 16, 2013 at 2:45 PM, Brian Kinlan (NOAA Affiliate) <brian.kinlan@noaa.gov> wrote:

-------- Original Message --------
Subject: Re: Fwd: geo-spatial statisitcs
Date: Wed, 16 Jan 2013 14:45:54 -0500
From: Brian Kinlan (NOAA Affiliate) <Brian.Kinlan@noaa.gov>
To: Randy Clark - NOAA Federal <randy.clark@noaa.gov>
CC: Charles Menza - NOAA Federal <Charles.Menza@noaa.gov>, Chris Caldow - NOAA Federal <chris.caldow@noaa.gov>

Randy,

Yes, this sounds very similar to the modeling problem we've tackled with seabirds and groundfish. The only problem is that it's not trivial to implement the solution. She probably needs to use some kind of geostatistical modeling - kriging - if she wants to account for the uneven density of sample points. Depending on the size and resolution of the regional model grid, she will probably also need to implement some kind of trend modeling (using lat/long/depth and perhaps additional environmental predictors). Trend modeling methods could range from GLM to GAM and boosted regression trees. One key element in the decision re: trend model is whether or not she needs maps of uncertainty. If she doesn't need uncertainty maps, she can pick from any one of a wide variety of algorithms to accomplish the trend modeling. However, if assessing uncertainty in a spatially explicit way (i.e., making an uncertainty map) is important, then only a small subset of trend modeling algorithms will work. To deal with the skewed/zero inflated distribution, she will most likely have to pursue some kind of hurdle or zero-inflated model. Finally, she will need to be familiar with cross-validation and bootstrapping concepts if it is important for her to assess the overall performance and accuracy of models.

For more information on our approach to this problem, she could explore chapter 6 of this tech memo (particularly appendix 6A):
http://ccma.nos.noaa.gov/ecosystems/coastalocean/ny_spatialplanning.aspx#products

For an alternative approach using Bagged Decision Trees, she could refer to the attached paper. However, this paper avoids dealing with spatial autocorrelation by simply coarsely binning data into 3km bins, which might not work or be acceptable for her purposes.

That said, it's important to realize that this is a complex statistical problem and requires a trained statistician to address. It would be unwise for someone without masters or phd-level training/experience in spatial statistics to attempt this kind of modeling effort. There is no out of the box solution that handles spatial autocorrelation/uneven sampling density and zero-inflation, so everything would have to be coded in R, Matlab, and/or SAS scripts. Depending on her background, she may want to rethink the requirements of the analysis or enlist someone with appropriate background as a collaborator. Although Charlie as a Fed could help her out if he had time, as a contractor I can't devote the time to help unless they have project funds they could BOP to us.

Feel free to relay this info to her...Charlie may also want to chime in.

cheers,
Brian

On 1/16/2013 1:05 PM, Randy Clark - NOAA Federal wrote:
> Hey guys, just got this request below. I immediately thought you two guys would be able to help? Start at the bottom, the request is from Laurel Smith, I don't have any contact info other than an email., but David Chevrier is at the NEFSC. Anyway, let me know if you can give Laurel and buzz and help out. Thanks!
> rc
>
> ---------- Forwarded message ----------
> From: Carlos Rivero - NOAA Federal <carlos.rivero@noaa.gov>
> Date: Wed, Jan 16, 2013 at 12:52 PM
> Subject: Fwd: geo-spatial statisitcs
> To: Randy Clark - NOAA Federal <randy.clark@noaa.gov>
>
> Hey Randy,
> I think this may up your alley. Can you pass this along to anyone that may be able to help?
> Thanks,
> Carlos
> ---------- Forwarded message ----------
> From: David Chevrier - NOAA Federal <david.chevrier@noaa.gov>
> Date: Tue, Jan 15, 2013 at 11:29 AM
> Subject: Fwd: geo-spatial statisitcs
> To: _NMFS GIS <nmfs.gis@noaa.gov>
>
>
> Hi,
> I'm sending this out on behalf of one of my users. They are looking for a best practice way to create some areas based on a detailed statistical analysis. Any help you could provide her would be greatly appreciated.
> Thanks,
> -dave
> ---------- Forwarded message ----------
> From: Laurel Smith - NOAA Federal <laurel.smith@noaa.gov>
> Date: Tue, Jan 15, 2013 at 11:04 AM
> Subject: geo-spatial statisitcs
> To: David Chevrier - NOAA Federal <david.chevrier@noaa.gov>
>
>
> Hi Dave,
>

> Thanks for passing this along, it would be a huge help if we could find a contact to consult with for statistical questions on ArcGIS tools.
>
> We are using randomly-distributed, stratified survey data to determine clusters of the densest biomass for each species. Possible data problems include areas of lower and higher sampling frequency (some areas with no samples and others with many samples taken closely together) that we don't want to effect the hot-spot density analysis, and non-normally distributed data that is skewed towards zero-biomass samples (tows that did not include that species). Once we have the density hot-spots identified, we need to aggregate ~30 layers of species data and determine areas of hot-spot overlap. These layers will need to be weighted based on several different ranked criteria.
>
> Thanks for your help with this,
> Laurel
>
> --
> Carlos Rivero
> Physical Scientist
> NOAA Fisheries
> carlos.rivero@noaa.gov
> 305.361.4484
> www.nmfs.noaa.gov
>
> "The contents of this message are mine personally and do not necessarily reflect any position of the Government or the National Oceanic and Atmospheric Administration."
>
> --
> Randy Clark
> Marine Biologist
> NOAA NCDDC/NCCOS Biogeography Branch
> 1021 Balch Blvd, Suite 1003
> Stennis Space Center, MS 39529
>
> 228-688-3732
> http://ccma.nos.noaa.gov/about/biogeography

--

--
****
Brian P. Kinlan, Ph.D

Marine Spatial Ecologist

NOAA National Ocean Service
Contractor, CSS Inc.
NCCOS-CCMA-Biogeography Branch
1305 East-West Hwy, SSMC-4, N/SCI-1, #9224
Silver Spring, MD 20910-3281

301.713.3028 x-157 (Tel)
301.713.4384 (Fax)
http://ccma.nos.noaa.gov/about/biogeography/
****