Use Case: Testing Relationships by Integrating Disparate Data Sources

Steven Worley with Tim Hoar, Rich Loft, Seth McGinnis, Don Middleton, Eric Nienhouse, Doug Schuster, Henry Tufo, Matthew Woitaszek

National Center for Atmospheric Research

Hector is an established Latin American scientist and ornithological expert on endemic bird populations of the regional cloud forests. With over two decades of field observation work he has an in depth understanding of the bird's annual and seasonal cycles, mating habits, and the dominant food sources they rely on. Over the past 8-10 years unexplained anomalies have appeared in his ongoing data time series. The question is "Why?" A few candidate factors are pressure from invasive species, recent anomalous weather patterns affecting food growth and availability, human encroachment on nesting areas, pollution from industrial development, or perhaps long-term climate shifts. Since causation in ecology is often multifactorial he knows that all or some combination of these may be to blame.

Hector has a research grant from the Tourism Board to evaluate the issues, because they are concerned about the impacts on the country's ecotourism business that relies in part on people being able to view wildlife. Since the grant is small he knows he must work with the limited resources at his disposal, and finding data that have free and open access is a critical enabling factor. The data that Hector has previously compiled are in simple tables that structure observations in time series, and he feels correlative studies are his best first order investigative tool, because they are simple, and with his spreadsheet program he can compute single and multi-variable statistical correlations.

Research activities:

- Hector approaches the weather and climate questions by first finding an atmospheric data portal that offers access to long-term time series. He begins with a generic search for "rain" and the query response returns a list of datasets that feature "precipitation" as part of general set of weather station data that also include temperature, wind, and relative humidity.
- Handy maps and calendar tools easily allow Hector to reduce a global view to a regional one and maintain the maximum time range on several of the datasets whose summaries are most interesting. He finds only a few national weather stations in the region, but is offered an option to extract, from the global archive, these local stations as a time-series, in a tabular format.
- There are some problems with the available data. The station locations are not near the bird nesting areas and they are at much lower elevations. Also, the precipitation data are very sparse, with many months of missing data scattered over various years. In contrast, the relative humidity and wind time series have no missing data for the past 30 years.

- Undeterred, he wonders if there might be a relationship between the short-term temperature and cloudiness codes (i.e. types of clouds and fog) that he has kept in his own records from the bird nesting areas, and the temperature and relative humidity at the weather stations.
- Correlation doesn't exist between the temperature records, but there is some correlation between relative humidity at the weather stations and low clouds and fog in his data. This correlation is stronger in some seasons when winds are in the onshore direction. Hector begins to feel that the weather station data could be a useful climate proxy for the bird study.
- In a seasonal visual comparison, the larger amounts of fig fruit in the cloud forest trees seem to align best with driest December-January-February season. However, the statistical correlation is insignificant, possibly because the records were too short, there are too many data voids in the precipitation record, or the sampling locations are were not geographically close enough.
- Recalling that the climate data portal recommended a related dataset for something called reanalysis – a term, incidentally unfamiliar to Hector - and associated environmental parameters pointing to river flow: he revisits the climate data portal.
- In the portal he finds a Spanish-language reanalysis tutorial, and discovers relative humidity is one variable that is available, but as gridded fields – something he has never worked with. Several encouraging things are discovered however: the reanalysis is 50 years long, the sampling is void-free at 6 hour intervals, there is an option to select a reduced region from the globe, and the gridded data can be interpolated to any set of points desired by the user.
- For Hector the tutorial also raises some cautionary awareness, because it discusses the limitations of reanalyses outputs. Data validity is reduced in areas where sparse input data are ingested into the assimilation model and coarse model resolutions can be inadequate for resolving local atmospheric conditions in regions that have steep topography and orographically forced weather phenomenon. This is critical data application information, because his study area is mountainous with frequent small-scale fog conditions, so reanalysis data has a potential to be misleading. He decides to proceed with caution, but bookmarks this information knowing that it should be discussed if reanalyses are used in any publication from this work.
- Hector has the choice to download the data or not: this is good, as his 5-year old computer's disk space is groaning under the load of this project as it is. Luckily, he discovers that the "gridded data can remain on temporary free storage in the analysis cloud for 30 days". He decides to get both the gridded and interpolated point data he needs and is required to select a format for the gridded data while the point data can be extracted in a format similar to the one he is already working with. He downloads the very small point data to his computer and leaves the gridded data on the cloud.
- Not too sure how to work with gridded data, he notes a menu that purports "Ways to view gridded data". Here there is a browser plugin that will display the data formats used to create his gridded subset. This is unfamiliar

territory for Hector, but how difficult could a browser plugin be - he has installed these before to access TV video news stories on the Internet.

- Back at the portal he ponders how river flow data is associated to rain, and realizes the integrative connection. Following the link he notices that he is at a different portal, but all his regional and temporal constraints from the climate portal have populated an interface to help him find the relevant river flow data. Maps show him his local data region has excluded a major river, so he expands the geographic dimension to include it. He requests the data in a time series tabular format – his data format "comfort zone". Immediately after submitting the request a query box pops up saying "re-gather gridded reanalysis data for this larger region?". Wow! That sounds like the right idea, because from a scientific perspective, it is a good methodology to keep the spatial and temporal domains synchronized where possible.
- He is now looking at data maps, from the gridded source in his browser, and time series line charts of various spreadsheet tables displayed as part of the spreadsheet application. Visually he is beginning to realize how the climate system works in this area, but he really wants the displays to work together!
- From the browser display plugin, he notices a help/more features tab. *Surprisingly, there are numerous features that can be added into and along side the gridded display - all in the same browser window, including animation.* He checks some introductory material and finds that the times series table data are easily connected to the display, and in fact time reference of both the gridded data and time series tables can be shared.
- The location of the weather station(s) and his records from the National Park can be spotted on the map, and a mark point marches along the time series line as the maps are animated in time. He is now visualizing together the original weather station data, the analyzed weather (reanalysis output), and river flow.
- Realizing the power of this tool he quickly creates time series spreadsheets of his point estimates of bird mating, hatch counts, and fledging success, and adds them to the display. Having the day-to-day precipitation plotted in an equivalent reference frame with the biological data immediately gives Hector ideas about how important weather extremes might be, whereas averaged (smoothed) data disguise short-term impacts.
- He is reaching the end of the day and knows that tomorrow he will keep going by pursuing other data about land use changes, industrial developments in the area, and tourist visits.
- How can he save all that he has achieved and restart later? He returns to the browser help/more features tab and finds "curate your work". The instructions seem simple and there is a small cost involved (based on the amount of data, and the amount of time (which is extendable) the project is preserved. For about 10% of his small Tourism Board budget he can preserve his work for 180 days. Why not do it? The result is surprising and impressive. The curation process pools <u>all</u> of his distributed data; time stamps everything, and puts it in a cloud location. It also archives all the settings of his current browser workspace along with software ID tags. It

even shows how the recovery will work by recommending he open a new browser window and click on a URL dedicated to him, this project, and this day. He tests it and it works. He heads home delighted with the progress, and eager to return and continue his research in the morning.