Workshop on Visualisation in Environmental Sciences (EnvirVis) (2013)
O. Kolditz, K. Rink and G. Scheuermann (Editors)

Visual Analytics for Exploring Changes in Biodiversity

A. Slingsby[1] and E. van Loon[2]

[1]School of Informatics, City University London, UK
[2]IBED, University of Amsterdam, The Netherlands

Abstract

We report on ongoing work in which we are designing a visual interface to a large database of species observation data. Our design allows the data to be explored and visually summarised by space, time and species, helping assess the data's suitability for helping answer questions about biodiversity. Key issues we are addressing include working with large datasets, dealing with varying spatial and temporal precisions and dealing with different qualities of data collection and sampling strategies. Our visual interface design is being informed by a set of research questions and a planned user-centred workshop.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—User-centered design

1. Introduction

The importance of maintaining long-term ecological records has been long recognised [MBB*10]. Since much of environment is now heavily influenced and managed by humans, monitoring biodiversity is increasingly important. Research questions generated in conservation biology [SAA*09] often require data that originate from different sources that differ in extent, observation properties and attributes. There has been rapid increase in availability of such data.

The Netherlands has an extensive network for collecting data on spatial and temporal distributions of flora and fauna from a variety of organisations, but not in a consistent form. The recently-established National Database of Flora and Fauna (NDFF) has collated these in semantically-consistent way. Over the past few decades, several *de jure* (Table 1 in [VvRS*12] and [JSRB06]) and *de facto* (Darwin Core and Access to Biological Collections Data) standards have emerged and are widely supported. The NDFF is comparable with Darwin Core with the additional advantage that it has not only syntactically but also semantically integrates datasets of different origins [VvRS*12]. It contains 40 million observations of 7000 species of mammals, birds, reptiles, amphibians, fish, invertebrates, plants and fungi [VvRS*12]. It offers good potential for studying species distributions and changes in biodiversity over space and time. Geographical and temporal visualisation techniques are widely used for exploring observation data [SWI*09], species distributions and species distribution models [FLF*11] that use vegetation and other data to help improve the precision of such mapping [SP03]. However, significant sampling and bias issues exist in the data due to the diversity of observations included.

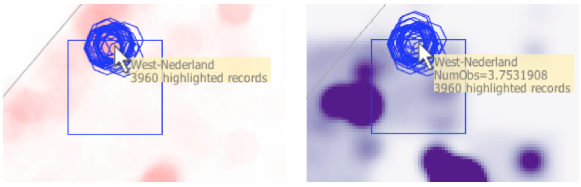

Figure 1: *Observation footprint map (left; blue footprint outlines) and density estimation map (right), indicating the extents and numbers of observations at the mouse cursor.*

In this short paper, we present, describe and reflect on an interactive visualisation design for exploring these data to help ecologists explore these biases and assess the extent to which they can help answer questions about biodiversity. The design draws on visual analytics techniques [TC05]. Techniques from Geographical Information Systems (GIS) are relevant, but often emphasise static cartography over exploratory interfaces and non-geographical aspects of data [Fis00]. We focus on the visualisation challenges of provid-

DOI: 10.2312/PE.EnvirVis.EnvirVis13.071-075

ing graphical interactive interfaces to raw observation data over long timescales with variation in spatial and temporal precision, different survey types, different abundance measures and large data sizes. We use visual analytics to study these and other issues (e.g. [Bow00]) related to data quality and suitability for studying biodiversity. We describe and justify our design decisions, discuss shortcomings that we are addressing and outline our plans for assessing how effective these techniques are for answering research questions from a group of ecologists.

2. Research questions

Biodiversity and conservation programmes require effective communication between scientists and resource managers [CVC12]. This involves a good understanding of the strengths and weaknesses of available data by all parties. Answers to simple questions concerning data availability, occurrence of species-groups over space and time, need to be readily available. Examples of such questions include:

- What is the coverage in space and time of observations on a given species group?
- Has key-species X been observed in region A?
- Is there a temporal trend in abundance of species Y?
- What is the trend in species richness (considering species group Z) in conservation area B?
- Are there areas where diversity and abundance of farmland birds have not declined over the past 30 years?

3. Data

Each of the 40 million observations in NDFF [VvRS*12] has a species or subspecies, a measure of abundance, a time and a place, measured at different precisions. Abundance may be recorded as a simple count, an estimate within upper and lower bounds or binary absence/presence. Some observations originated from systematic surveys in which absence was specifically recorded, but most do not record absence. Observations have time recorded at various different precisions (decade, year, month and day), modelled as the time window within which the observation was recorded. This is also the case for space, which NDFF models as a polygon within which the observation was recorded, including 5km squares, 1km squares and circular regions that represent point locations recorded at low precision.

NDFF also provides a hierarchical species taxonomy which includes information about whether they feature on a variety of species protection lists. Measuring biodiversity in terms of administrative boundaries is important for official statistics and compliance reasons, so we incorporate a hierarchy of geographical regions into our visual analytics.

4. Design

Data is read from a flat file produced by a single SQL query, but could easily read the data directly from the database. Although we focus on visual and interactive design, the technical challenges of dealing with large datasets is significant. The prototype can deal with around 6 million records interactively. In common other visual analytics designs for spatiotemporal data and questions, we use coordinated map-based and timeline-based views. We also added a hierarchical view representing species. A full screenshot can be seen in Fig. 6.

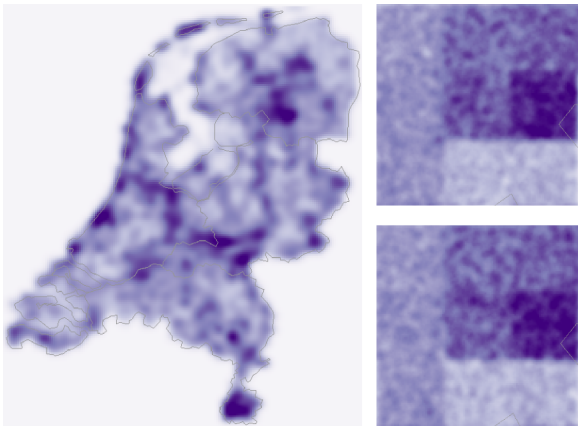

Figure 2: *Density of the 5 million observations made within 5km^2 grid squares; for the whole country (left) and zoomed to a 1.5km×1.5km area (right).*

4.1. Map view

The zoomable and pannable map view uses the Dutch National Grid to show where the observations were taken, depicted as a footprint, density or choropleth map. Parameters of the display including colour scaling and changing kernel size can be changed interactively.

The zoomed-in portion of the footprint map in Fig. 1 (left) shows whole extents of the observation footprints. Where more observations were made, the red colour is darker. Brushing by holding down SHIFT and moving the mouse of the map identifies the observation extents at the mouse cursor and the number of observations there, details that cannot be seen on the full footprint map.

The footprint map has the effect of giving greater saliency to the least spatially precise observations. To address this, we use a density estimation surface in Fig. 1 (right) in which each observation has been allocated a random location within its footprint. On-the-fly density estimation surfaces are generated using a resizable Cressman filter [COO92] which adapts to the scale at which the map is being viewed. A different random position is used on each map redraw, so that the visual stability between map redraws indicates the effect of the spatial precision at the zoom level used. As an illustration, in Fig. 2 we are only showing observations (∼5

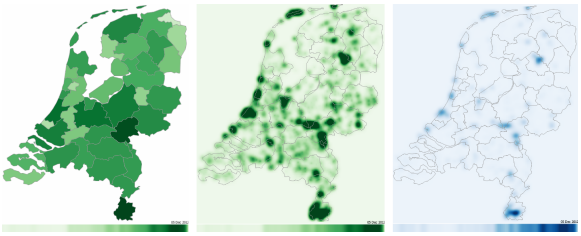

Figure 3: *Choropleth map (NUTS3 administrative regions) of the number of species (left), density map of the number of species (centre) and density map of the number of protected species (right) for a random sample of 1 million records. Timelines are displayed below.*

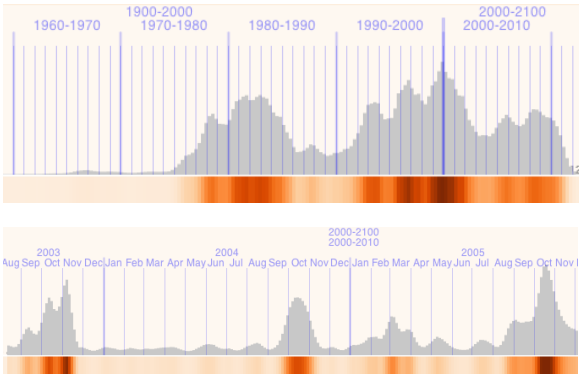

Figure 4: *Average abundance of all species (random sample of a million observations) on a timeline, using zooming to investigate different temporal scales. Moving the mouse over this adds a histogram and time labels.*

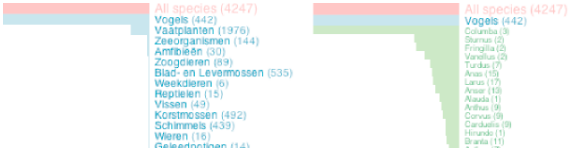

Figure 5: *Colour distinguishes between genus (left) and species (right). Barcharts indicate the number of subspecies within each genus/species.*

million) made within 5km^2 grid squares. The differences between in the zoomed-in portions, are due to different random positions. If we assume observations are equally likely to be anywhere in the footprint, there are enough observations for the effect of the low spatial precision to be negligible when considering overall density, even through the spatial precision is very coarse for the zoom level.

Choropleth maps are important in an administrative context, as biodiversity measures by administrative units are sought. We provide a variety of geographies including the European standard NUTS1/2/3 regions, as well as a number of Dutch-specific regions at finer spatial resolutions. From data exploration and biodiversity perspectives, these are the least useful views because of the Modifiable Areal Unit Problem (MAUP) [Ope83] which is reduced in the density maps by the smoothing kernel and the finer spatial units. MAUP is conflated by the additional instability caused by the random position allocation of the less spatially precise observations, however, as with the density maps, the effect of this can be determined visually by redrawing the map multiple times.

In addition to the distribution of observations, we also map average abundance (orange), unique species count (green) and unique protected species count (blue), using sequential ColorBrewer [Bre09] schemes of different hues, consistently.

4.2. Timeline view

The timeline view is implemented as a one-dimensional version of the map view. As with the spatial information, observations are made during imprecise temporal windows, often by decade or year. Brushing the timeline highlights the temporal windows of observations made during that time. As with the map, random perturbation on each timeline redraw is employed. Zooming allows the data to be studied at different temporal scales (Fig. 4), colour depicts the same values as shown on the map and a one-dimensional version of the resizable Cressman Filter smooths the values by the user defined amount. Timelines are placed below the maps (Fig. 3), but when the mouse is moved over the timeline a histogram that double-encodes the values and time labels are added (Fig. 4). The histogram's use of aligned length provides a better basis for estimating and comparing values than colour lightness and the zooming combined with the Cressman kernel enables cycles in decadal, annual and monthly species abundance to be determined.

4.3. Species view

A barchart of the number of subspecies contained within each genus and species accompany the species list, which can be expanded as in Fig. 5. Different colours distinguish genus, species and subspecies (red indicates a 'protected' status) and they can be ordered by subspecies frequency and in alphabetical order. This zoomable list enables overviews of the hierarchical structure of species taxonomy.

4.4. Selection

As well as providing spatial, temporal and species overviews, each view provides a basis for selection. Fig. 6 shows a spatial selection (the species list has been correspondingly updated showing that only bird and reptile obser-

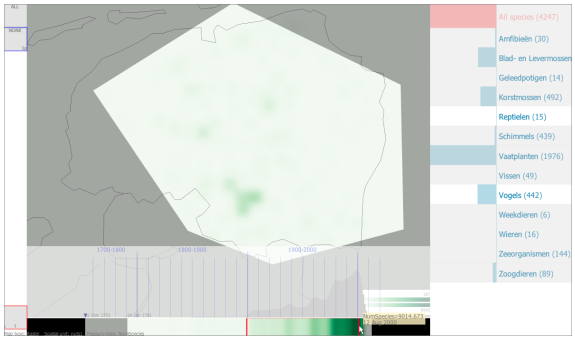

Figure 6: *Screenshot from prototype application. Spatial selection and species selections have been applied. A temporal selection (for the past century) is in progress. For this region, the unique number of species observed has been increasing (due to more observations being made over time).*

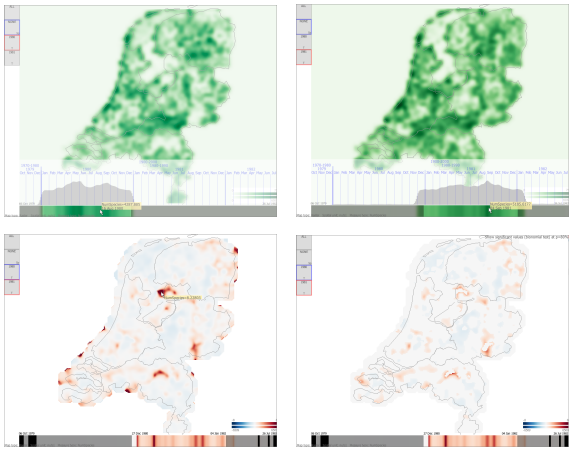

Figure 7: *Two selections of observations 1980 and 1981 have been saved and can be flipped between (top). Setting 1980 as the baseline, red indicates a higher proportional difference; blue indicates lower (bottom left). On the basis of a binomial statistical test, values that are <80% statistically significant have been removes (bottom right).*

vations exist in the spatial selection) and a temporal selection for the past century being selected (red outline).

Spatial, temporal and species aspects of selections can be saved, named and compared (see next section) as in [SDW11]. These appear on the left of the screen (Fig. 6 has the default 'all' and 'none'; Fig. 7 has two more).

4.5. Comparison

The zoomable map, timeline and spatial views along with the selections they facilitate, show spatial, temporal and species distributions of subsets of the data; e.g. the seasonal cycles in species abundance in Fig. 4. This is a form of simple comparison, but does not let us easily look at spatial distributions over time or temporal distributions over space.

Saving selections and then being able to apply these on-demand, allows quick switching between views in Fig. 7 (top), in which 1981 appears to have a higher overall number of species. Such comparison is difficult. To address this, we encode the difference between these directly in Fig. 7 (bottom left) by setting the 1980 selection as a baseline. Red indicates that the value was higher in 1981 than 1980. These red areas are stable on map redraws. The timeline is dominated by red, indicating that unlike spatially, the species number is consistently higher throughout the year. Using the bionomial proportions statistical test [DeV08] and only colouring parts of the map that are statistically significant (at an 80% p-value) in Fig. 7 (bottom right), there are geographical areas where this increase is statistically significant.

5. Addressing research questions and ongoing work

The graphical representation of spatial and temporal distributions of species, the ability to constrain, save and recall selections by space, time and species and the ability to have results reported as number of observations, average abundance or numbers of species present, enables answers to be obtained to the research questions listed in section 2. A planned workshop with ecologists will help evaluate the success of the design, how initiative the user interface is and whether participants can be confident in the results being reported.

Sampling bias in observation data often requires the use of species distribution models [SP03]. We hope to help establish the degree with which this is true for NDFF – as an example of a large, comprehensive and semantically-rich dataset – using exploratory interactive visualisation. We hope that our findings will help inform the design of similar databases in future. Sampling bias is perhaps the most significant issue in this work. By identifying systematic surveys with species lists in which observation absence implies absence, we hope to explore such sampling bias.

6. Conclusion

The prototype that implements our visual and interactive design is able to work with reasonably large subsets of the database interactively and methods for exploring spatial, temporal and species aspects of the data are able to address the research questions to some degree. Methods for dealing with the imprecise spatial and temporal precision of observations and for exploring and querying the data at different spatial and temporal scales work well. These can be used to address simple research questions that relate to spatial and temporal distributions. We think this demonstrates the validity of this approach and we are hopeful that our planned ongoing work will help ecologists better exploit this new semantically-rich and consistent database of biodiversity.

References

[Bow00] BOWKER G. C.: Mapping biodiversity. *International Journal of Geographical Information Science 14*, 8 (2000), 739–754. URL: http://www.tandfonline.com/doi/abs/10.1080/136588100750022769. 2

[Bre09] BREWER C.: Colour advice for cartography. http://www.colorbrewer.org, 2009. 3

[COO92] COOPER W.: Cressman analysis. http://www.asp.ucar.edu/colloquium/1992/notes/part1/node119.html, 1992. 2

[CVC12] CAUDRON A., VIGIER L., CHAMPIGNEULLE A.: Developing collaborative research to improve effectiveness in biodiversity conservation practice. *Journal of Applied Ecology 49*, 4 (2012), 753–757. URL: http://dx.doi.org/10.1111/j.1365-2664.2012.02115.x. 2

[DeV08] DEVORE J.: *Probability and Statistics for Engineering and the Sciences: Enhanced [With Glossary of Symbols Booklet]*. Available 2010 Titles Enhanced Web Assign Series. Brooks/Cole, Cengage Learning, 2008. URL: http://books.google.co.uk/books?id=Wbym40WgsXMC. 4

[Fis00] FISHER P. F.: Is GIS hidebound by the legacy of cartography? *Cartographic Journal, The 35*, 1 (1998-06-01T00:00:00), 5–9. URL: http://www.ingentaconnect.com/content/maney/caj/1998/00000035/00000001/art00002. 1

[FLF*11] FERREIRA N., LINS L., FINK D., KELLING S., WOOD C., FREIRE J., SILVA C.: Birdvis: Visualizing and understanding bird populations. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2374–2383. doi:10.1109/TVCG.2011.176. 1

[JSRB06] JONES M. B., SCHILDHAUER M. P., REICHMAN O., BOWERS S.: The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics 37*, 1 (2006), 519–544. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.37.091305.110031. 1

[MBB*10] MAGURRAN A. E., BAILLIE S. R., BUCKLAND S. T., DICK J. M., ELSTON D. A., SCOTT E. M., SMITH R. I., SOMERFIELD P. J., WATT A. D.: Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution 25*, 10 (2010), 574 – 582. doi:10.1016/j.tree.2010.06.016. 1

[Ope83] OPENSHAW S.: *The modifiable areal unit problem*, vol. 38. Geo Books Norwich, 1983. 3

[SAA*09] SUTHERLAND W. J., ADAMS W. M., ARONSON R. B., AVELING R., BLACKBURN T. M., BROAD S., CEBALLOS G., CÔTÉ I. M., COWLING R. M., DA FONSECA G. A. B., DINERSTEIN E., FERRARO P. J., FLEISHMAN E., GASCON C., HUNTER JR. M., HUTTON J., KAREIVA P., KURIA A., MACDONALD D. W., MACKINNON K., MADGWICK F. J., MASCIA M. B., MCNEELY J., MILNER-GULLAND E. J., MOON S., MORLEY C. G., NELSON S., OSBORN D., PAI M., PARSONS E. C. M., PECK L. S., POSSINGHAM H., PRIOR S. V., PULLIN A. S., RANDS M. R. W., RANGANATHAN J., REDFORD K. H., RODRIGUEZ J. P., SEYMOUR F., SOBEL J., SODHI N. S., STOTT A., VANCE-BORLAND K., WATKINSON A. R.: One hundred questions of importance to the conservation of global biological diversity. *Conservation Biology 23*, 3 (2009), 557–567. URL: http://dx.doi.org/10.1111/j.1523-1739.2009.01212.x. 1

[SDW11] SLINGSBY A., DYKES J., WOOD J.: Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2545–2554. URL: http://openaccess.city.ac.uk/437/. 4

[SP03] STOCKWELL D., PETERSON A.: Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. *Ecological Indicators 3*, 3 (2003), 213 – 221. URL: http://www.sciencedirect.com/science/article/pii/S1470160X03000451. 1, 4

[SWI*09] SULLIVAN B. L., WOOD C. L., ILIFF M. J., BONNEY R. E., FINK D., KELLING S.: ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation 142*, 10 (2009), 2282 – 2292. URL: http://www.sciencedirect.com/science/article/pii/S000632070900216X. 1

[TC05] THOMAS J. J., COOK K. A.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0769523234. 1

[VvRS*12] VEEN L., VAN REENEN G., SLUITER F., VAN LOON E., BOUTEN W.: A semantically integrated, user-friendly data model for species observation data. *Ecological Informatics 8*, 0 (2012), 1 – 9. URL: http://www.sciencedirect.com/science/article/pii/S1574954111000926. 1, 2