Summer Course

Next Generation Data Management in Movement Ecology

San Michele all'Adige, Trento, Italy

SUMMARY OF TALKS, STUDY CASES AND EXERCISES

Module I

DAY 1 (Wednesday, 01-07-2015)

LESSON 1: Introduction to SQL – Antonio Galea

This lesson briefly introduces the concept of a database management system, and then proceeds to explore the basic SQL statements needed to retrieve data from a single database table. You will see how information from different tables can be combined in a query. At the end of the lesson, you will be able to get data from the database specifying criteria to retrieve the desired subset of records.

DAY 2 (Thursday, 02-07-2015)

LESSON 2: Installing PostgreSQL and data manipulation – Antonio Galea

In this lesson you will install a PostgreSQL server, practice with its graphical front-end pgAdmin, and learn how to create the structure and manipulate the data. You will also get to know about basic database administration tasks, like user management and backup/restore procedures.

DAY 3 (Friday, 03-07-2015)

LESSON 3: Introduction to Spatial Databases – Antonio Galea & Federico Ossi

Our last lesson will focus on introducing spatial data types, which will enable you to accomplish richer analysis of wildlife tracking data. By the end of the day you will be able to juggle with geometry columns, make the database compute for you answers to spatial questions, import vector data. Moreover, it will be shown a basic application of PostgreSQL and PostGIS for ecological investigation.

Module II

DAY 1 (Monday, 06-07-2015)

TALK: Introduction to wildlife tracking – Federico Ossi

The ideal objective of any movement ecology study is rooted in relevant ecological questions that can contribute to theory, and inform conservation and management actions. For example, a study on natal dispersal can help in identifying barriers to gene flow; or an analysis on environmental characteristics affecting population performance can support decisions on protection areas and conservation corridors. Whatever the question, a necessary step should be the evaluation of the appropriate methodology, and specifically deciding whether individual marking devices such as GPS (or other sensors) are the most effective and informative approach to pursuing the final goal. If the answer is yes, the use of these tools, which is always increasing, should be paralleled by an equally rapid development of procedures to manage and integrate animal movement data sets. This potentially leaves a gap between the acquisition of data, and the overarching scientific questions these tools have the potential to address. What to do with these data? How to handle, manage, store and retrieve them, and how to eventually feed them to analysis tools such as statistics packages or Geographic Information Systems (GIS), and test scientific hypotheses? We will explore these questions, potential tools and perspectives, putting the roots for a deeper investigation within the "world" of database management systems for fully exploiting the collected data. This is the way to follow to fill the gap between the state-of-the-art knowledge on data management, and its application to wildlife tracking data.

TALK: An answer to our management needs: relational database – Ferdinando Urbano

In the last 15 years, new wildlife tracking and telemetry technologies have become available to scientists, leading to substantial growth in the volume of wildlife tracking data. In the future, one can expect an almost exponential increase in collected data as new sensors are integrated into current tracking systems. A crucial limitation for efficient use of telemetry data is a lack of infrastructure to collect, store and efficiently share the information. Large data sets generated by wildlife tracking equipment pose a number of challenges: to cope with this amount of data, a specific data management approach is needed, one designed to deal with data scalability, automatic data acquisition, long-term storage, efficient data retrieval, management of spatial and temporal information, multi-user support, and data sharing and dissemination. The state-of-the-art technology to these challenges is the relational DataBase Management System (DBMS), with its dedicated spatial extension. DBMS are efficient, industry-standard tools for storage, fast retrieval and manipulation of large data sets, as well as data dissemination to client programs or web interfaces. In the future we expect tools able to deal with both spatial and temporal dimensions of animal-movement data, such as spatio-temporal databases, and to scale to a even larger size of data sets of what is called Big Data era.

EXERCISE: Storing your tracking data in an advanced database platform: PostgreSQL – Ferdinando Urbano

The state-of-the-art technical tool for effectively and efficiently managing tracking data is the spatial relational database. Using databases to manage tracking data implies a considerable effort for those who are not already familiar with these tools, but this is necessary to be able to deal with the data coming from the new sensors. Moreover, the time spent to learn databases will be largely paid back with the time saved for the management and processing of the data. In this lesson, you are guided through how to set up a new database in which you will create a table to accommodate the test GPS data sets. You create a new table in a dedicated schema.

TALK: Extending the database with other meta information: capture, mortality, population data – Federico Ossi

Meta information on e.g. mortality rate, density, or capture success can importantly enrich the amount of available data for biological investigation, permitting ecologist to reconnect individual movement patterns with population dynamics. Relational databases offer a valid tool in this sense to collect information from one (or more) population under investigation. We discuss the different steps of population data management, exploring some potential fields of application.

EXERCISE: Managing and modelling information on animals and sensors - Ferdinando Urbano

GPS positions are used to describe animal movements and to derive a large set of information, for example about animals' behavior, social interactions, and environmental preferences. GPS data are related to (and must be integrated with) many other sources of information that together can be used to describe the complexity of movement ecology. In a database framework, this can only be achieved through proper database data modelling, which depends on a clear definition of the biological context of a study. Particularly, data modelling becomes a key step when database systems manage a large set of connected data sets that grow in size and complexity: it permits easy updates and modification and adaptation of the database structure to accommodate the changing goals, constraints, and spatial scales of studies. In this lesson, you will extend your database with two three new tables to integrate ancillary information useful to interpreting GPS data: one for GPS sensors, one for animals, and one for captures.

STUDY CASE: Eurodeer – Francesca Cagnacci

The EUropean ROe DEER Information System (EURODEER) is an open project to support a collaborative process of data sharing to produce better science. It is based on a spatial database that stores shared movement data on roe deer to investigate variation in roe deer behavioural ecology along environmental gradients or population responses to specific conditions, such as habitat changes, impact of human activities, different hunting regimes. EURODEER group is trying to fully explore the opportunities given by the new monitoring technologies for

conservation and management at both local and continental scale. The spatial database, built upon open source software (PostgreSQL + PostGIS + R) and hosted at Fondazione Edmund Mach, can be connected to a large set of client applications (GIS, web interfaces, statistics) to help storing, managing, accessing and analysing GPS data from several research groups throughout Europe. The use of EURODEER might permit to synthesize our knowledge of roe deer ecology into a wider and more complex picture, that would allow to clarify ecosystemic relationships (e.g., resource balance), reveal evolutionary patterns (e.g., animal performance), and underpin predictions on future scenarios (e.g., climate change effect).

DAY 2 (Tuesday, 07-07-2015)

EXERCISE: From data to information: associating locations to animals – Ferdinando Urbano

When position data are received from GPS sensors, they are not explicitly associated with any animal. Linking GPS data to animals is a key step in the data management process. This can be achieved using the information on the deployments of GPS sensors on animals (when sensors started and ceased to be deployed on the animals). In the case of a continuous data flow, the transformation of GPS positions into animal locations must be automated in order to have GPS data imported and processed in real-time. In this lesson, you extend the database with two new tables, gps_sensors_animals and gps_data_animals. As additional material, a set of dedicated database triggers and functions is presented that add tools to automatically manage the association of GPS positions with animals.

STUDY CASE: U.S. IOOS Data Management for Marine Animal Telemetry – Hassan Moustahfid

The growing volume of marine animal telemetry data holdings, the large diversity of tag types and data formats, and the general lack of data management are not only complicating integration and synthesis of animal telemetry and tracking data but potentially threatening the integrity and longer-term access to these valuable datasets. To address this critical gap, the US IOOS Animal Telemetry Data Management and Visualization System (ATN DAC) has been developed to provide an integrated system of most known transmitters and tracking systems. The ATN DAC leverages off several systems within the existing data product delivery network to provide data to the ATN DAC interface. The data delivery processes include direct PostreSQL database queries, accessing NOAA - Environmental Research Division's Data Access Protocol (ERDDAP) services, and extraction from automatically generated standard delimited data files. Those data are unified within the ATN DAC system then delivered to the ATN DAC interface. In this presentation I will provide an overview of the ATN DAC data management and visualizations capabilities and linking this data service to ocean models and applications.

EXERCISE: Spatial is not special: how to manage the locations data in a spatial database: PostGIS – Ferdinando Urbano

A wildlife tracking data management system must include the capability to explicitly deal with the spatial component of movement data. GPS tracking data are sets of spatio-temporal objects (locations) and the spatial component must be properly handled. You will now extend the database adding spatial functionalities through the PostgreSQL spatial extension PostGIS. PostGIS introduces the spatial data types (both vector and raster) and a large set of SQL spatial functions and tools, including spatial indexes. This possibility essentially allows you to build a GIS using the existing capabilities of relational databases. In this lesson, you will implement a system that automatically transforms the GPS coordinates generated by GPS sensors from a pair of numbers into spatial objects.

TALK: Geographical data in ecology: from local to global spatial scales – Duccio Rocchini,

How a geographical perspective might help understanding ecological processes? Facing this issue represents the core aim of this talk, which attempts to disentangle the role of geographical drivers to solve different ecological problems related to: biodiversity, species distribution modelling, sampling effort and uncertainty.

EXERCISE: Environmental layers: integration and management of spatial ancillary information – Ferdinando Urbano

Animals move in and interact with complex environments that can be characterized by a set of spatial layers containing environmental data. Spatial databases can manage these different data sets in a unified framework, defining spatial and non-spatial relationships that simplify the analysis of the interaction between animals and their habitat. This simplifies a large set of analyses that can be performed directly in the database with no need for dedicated GIS or statistical software. Such an approach moves the information content managed in the database from a geographical space to an animal's ecological space. This more comprehensive database model of the animals' movement ecology reduces the distance between physical reality and the way data are structured in the database, filling the semantic gap between the scientist's view of biological systems and its implementation in the information system. This lesson shows how vector and raster layers can be included in the database and how you can handle them using (spatial) SQL. The database built so far is extended with environmental ancillary data sets.

EXERCISE: How to extract environmental information related to location data - Ferdinando Urbano

The association of GPS position with environmental attributes can be part of the preliminary processing before data analysis where a set of procedures is created to intersect ancillary layers with GPS positions. Database tools like triggers and functions can be used for this scope. The result is that positions are transformed from a simple pair of numbers (coordinates) to complex multi-dimensional (spatial) objects that define the individual and its habitat in time and space,

including their interactions and dependencies. In an additional step, position data can also be joined to activity data to define an even more complete picture of the animal's behavior. Once this is implemented in the database framework, scientists and wildlife managers can deal with data in the same way they model the object of their study as they can start their analyses from objects that represent the animals in their habitat (which previously was the result of a long and complex process). Moreover, users can directly query these objects using a simple and powerful language (SQL) that is close to their natural language. All these elements strengthen the opportunity provided by GPS data to move from mainly testing statistical hypotheses to focusing on biological hypotheses. Scientists can store, access, and manipulate their data in a simple and quick way, which allows them to formulate biological questions that previously were almost impossible to answer for technical reasons. In this lesson, GPS data and ancillary information are connected with automated procedures. In an extra section, an example is illustrated to manage time series of environmental layers that can introduce temporal variability in habitat conditions.

TALK: Finding our way on the sharing and re-use of animal telemetry data in Australasia – Hamish Campbell

There are a growing number of online repositories for the storage, sharing, and reuse of animal telemetry data. The utopian idea of researchers sharing their hard won data for the good of the community is certainly valid, but in practice fraught with complexity. Here I will discuss how we are dealing with these issues in that far-flung corner of the globe, Australasia.

STUDY CASE: ZoaTrack- A web-based platform for assessing species movement and occupancy – Hamish Campbell

The zoaTrack platform is an online facility for the calculation of space use from time-series positional fixes. It performs a range of data cleansing, filtering and home-range density estimation functions, and then creates visual outputs of these analyses in a POSTGIS environment. Users can overlay and compare a range of home-range calculations with environmental layers, to assess animal movement, space use, and habitat association.

DAY 3 (Wednesday, 08-07-2015)

TALK: From locations to steps: the movement model – Mathieu Basille

This talk will present the fundamentals of movement ecology, a conceptual framework designed to address basic and applied questions related to movement. Using the information embedded in temporal series of locations (i.e. a trajectory), we will see how we can refine habitat selection and home range approaches, by integrating a mechanistic decomposition of movement. We will explore a few fundamental papers from the movement ecology literature, and see what's coming next!

EXERCISE: The movement model: implementation – David Bucklin

In this exercise, we will implement the movement model by creating trajectories from animal relocations in PostGIS, and viewing them in QGIS. Students will calculate geometric parameters of movement (speed, angle of movement, etc.) and investigate how movement trajectories are related to environmental factors (e.g., land cover, elevation, and roads).

TALK: Sensors data: how to ingest into the database – Timothy Giles

Sensor Data: Ingesting in to a database, processes and considerations. We take a look at the history of importing (sensor) data, how it has changed over the years from importing to ingesting. A review of some of the processes we have followed with advice and critique of techniques and what the future brings us and how we are adapting to it.

STUDY CASE: WRAM – Timothy Giles

WRAM: What is it? The history of WRAM. Why and how has the WRAM system changed, why put your data in WRAM or any other system? What benefits can you leverage by using such a system? An overview of the topology and system architecture that delivers secure robust data storage.

EXERCISE: Consolidation and use of the spatial database built so far – Ferdinando Urbano

In the first 6 lessons, you created a database with tracking data and related information (animals, sensors, deployments). You defined a set of procedures to automatically manage the update of the database whenever new GPS data are acquired. You further extended the database including ancillary environmental information. In this lesson, you play with your data to apply the SQL skills learn so far (and particularly during the first part of the summer school) to tracking data. Here 9 exercises are proposed that simulate some possible data processing that can be done in the database framework. You are invited to try to solve them autonomously and then check the results during at the end of the lesson with all the other students.

DAY 4 (Thursday, 09-07-2015)

TALK: GPS technology and future perspectives – Nicola Gadow

Technical background of GPS technology and future perspectives GPS collars are one of the best possibilities to receive tracking data from wild animals. Beside of the GPS positions you can have much more information about your study animal like activity, temperature, mortality, heart rate etc. We will focus on integrated sensors (3-axis activity sensor (storing raw data), mortality sensors) and external sensors (proximity sensors, separation sensors / fawn collars, VIT, MIT, Ruminal Heart Rate Data Logger). Additionally I will explain the intelligent collar programming which enables the collar to switch the GPS schedule according to several events (e.g. exceeded activity threshold, virtual fence event, proximity event). Last point is to talk about future options

such as processing activity data together with GPS data (Noldus company) as well as the new snapshot technology which will come soon.

TALK: Accelerometers: investigating animal's activity – Anne Berger

Over the past two decades, an explosion of research on remote monitoring of animal behaviour using accelerometers has broken down the old limits of purely observational studies: Animal-attached acceleration sensors measure the change in velocity of the body or parts of the body over the time and can quantify fine-scale movements and body postures without animal reaction and independently from observer bias and visibility of the observed animal. Nowadays, accelerometers (or other sensors) are additionally integrated on commercial GPS-devices to interpret the behaviour of tagged animals. Although the measurement procedure is essentially identical in all different accelerometers that are used on animals, the data obtained from various GPS-devices differ widely due to internal data processing methods, resolution and sensitivity. The general structure of these data and an overview of possibilities for their analysis are given in the talk.

EXERCISE: Integrating activity data into the database – Ferdinando Urbano

In the previous lessons, you have exclusively worked with GPS position data. We showed how to organise these data in databases and how to link them to environmental data. In this lesson, we introduce an example of data recorded by another type of sensor: acceleration data, which can be measured by many tags where they are associated with the GPS sensors and are widely used to interpret the behaviour of tagged animals. In this exercise, you will learn how to integrate acceleration data into the database while in the theoretical part the main data management challenges for this kind of massive data sets are discussed.

EXERCISE: Working with activity data: some examples – Ferdinando Urbano & Anne Berger

Once acceleration data are stored into the database, they are ready to be analyzed. Acceleration data can provide valuable information and can help to create the so called semantic trajectories, i.e. description of the movement that identify specific actions of animals while they interact with the surrounding environment. In this lesson you will see some examples of analysis.

TALK: A new type of biologged data: spatial explicit contact detection – Federico Ossi & Francesca Cagnacci

One decade ago, proximity loggers, i.e. biologgers which permit to detect contacts within individuals wearing them, appeared for the first time in animal ecology world. From there on, these tools have been more and more used to address a variety of ecological issues, from disease dynamic transmission to competition patterns. We discuss strengths and weaknesses of these tools. We also present a recently developed proximity loggers which, for the first time to our knowledge, couples information on proximity patterns with spatial data collected through a triggered GPS acquisition system.

EXERCISE: Data quality: how to manage wrong locations – Ferdinando Urbano

Tracking data can potentially be affected by a large set of errors in different steps of data acquisition and processing. Erroneous data can heavily affect analysis, leading to biased inference and misleading wildlife management/conservation suggestions. Data quality assessment is therefore a key step in data management. In this lesson, we especially deal with biased locations, or 'outliers'. While in some cases incorrect data are evident, in many situations it is not possible to clearly identify locations as outliers because although they are suspicious (e.g. long distances covered by animals in a short time or repeated extreme values), they might still be correct, leaving a margin of uncertainty that often depends on the specific analysis. In this lesson, different potential errors are identified and a general approach to managing outliers is proposed that tags record rather than deleting them.

EXERCISE: Analyzing and managing movement data: representations, methods, and tools in the database framework – Ferdinando Urbano

The objects of movement ecology studies are animals whose movements are usually sampled at more-or-less regular intervals. This spatio-temporal sequence of locations is the basic, measured information that is stored in the database. Starting from this data set, animal movements can be analysed (and visualized) using a large set of different methods and approaches. These include (but are not limited to) trajectories, raster surfaces of probability density, points, (home range) polygons, and tabular statistics. Each of these methods is a different representation of the original data set that takes into account specific aspects of the animals' movement. The database can help to implement these multiple representations of tracking data. In this lesson, some examples of methods for implementing GPS tracking data representations into a spatial database (i.e. with SQL code and database functions) are introduced.

TALK: Data sharing and dissemination in movement ecology: why and how to archive your data – Sarah Davidson

This talk covers options for archiving your research data and making it available for future re-use. Fifty years from now, where will your animal tracking data be? Collecting animal tracking data requires a huge amount of time and money, impacts the animals that carry the tags, and provides unique records about these animals that cannot be replicated. As data owners, you should consider how your data can be shared with others, archived, and made available for future re-use. In addition to making your data available to the scientific community and the public—which is increasingly required by funding providers and journals—personal benefits to sharing your data can include increased citations of your research and new collaborations that combine data to answer new questions. In organizing data for archiving, it is essential that the data and data set are thoroughly described, wherever possible using established vocabularies and metadata schema. This allows others to fully understand your data (e.g., to know units of measurement, collection methods) and to search for and access your data set (interoperability).

Three options for archiving a research data set are to (1) publish it in a data repository (e.g. Dryad, Movebank Data Repository), (2) publish a data paper (e.g. Biodiversity Data Journal, Ecological Archives, Scientific Data), or (3) add it to a shared database (e.g. DataONE, GBIF, KNB, animal tracking databases). Proper data publishing should include a review process, assignment of a persistent identifier such as a DOI, and licensing to define explicit conditions for re-use. Fortunately, the growing number of options for archiving movement ecology data sets make all of this quite easy, especially if you have a well-managed data set. Thus data sharing and archiving are a natural next step after building your own spatial database as taught in this course.

STUDY CASE: Movebank – Martin Storhas

Movebank (www.movebank.org) is a free, online infrastructure to help researchers store, manage, share, analyze, and archive animal tracking data. Movebank is hosted by the Max Planck Institute for Ornithology and is open to all researchers, species, study areas, and funding sources. Tracking data in Movebank are stored in a PostgreSQL database. Queries, visualization, and web tools are run via Java and a Java application server. Data are stored and accessed within user-created studies. Users import data to the Movebank database using near-live data feeds—available for a growing number of tag types—or by uploading data files and mapping dataset attributes to defined Movebank terms. Metadata ("reference data") can be added using defined terms to further describe the Tags, Animals, and Deployments. Within a study, data owners set flexible access permissions that allow them to keep parts or all of the study completely private, share them with other registered users, or make them available to the public. Users with access to data can apply filters, access it in R and other programs for analysis, and annotate hundreds of environmental variables from global remote sensing and weather data products using the Env-DATA System. To allow permanent public archiving, the Movebank Data Repository gives users the option to publish datasets associated with peer-reviewed papers and receive a DOI. With over 1,900 user-created studies, representing almost 500 taxa and containing over 195 million animal locations, Movebank offers unique opportunities for data sharing, outreach, and collaboration.

DAY 5 (Friday, 10-07-2015)

EXERCISE: There and back again – part I: Analyzing movement data in the R statistical environment – Mathieu Basille

In this exercise, we will move from the database to R, an open source programming language and environment for statistical computing and graphics, and specifically the package 'adehabitatLT' dedicated to the management and study of trajectories. We will first build a trajectory step by step, and then use graphical approaches to explore and understand it. More complex functions to clean and extend a trajectory will also be introduced, as well as a full-blown example of movement-based habitat selection analysis.

TALK: The Qualitative Trajectory Calculus (QTC) as a tool to analyse complex interactions between moving animals – Nico Van de Weghe

The Qualitative Trajectory Calculus (QTC) is a qualitative calculus to represent and reason about moving objects. In this talk, the basis of this calculus will be presented, followed by its potential applications within the area of animal movement.

EXERCISE: There and back again, part II: Connecting PostGIS and R – Mathieu Basille

In this exercise, we will connect R to the database to enable bidirectional transfers. We will use various R packages with different features and assets: 'RPostgreSQL' for basic data transfer, 'rgdal' for spatial objects, and 'rgeos' for WKT representations. In a second step, we will wrap the most common and useful SQL queries into R functions in the package 'rpostgis'.

EXERCISE: The movement model: recap and integration with the database concepts developed so far – Mathieu Basille & David Bucklin

In this exercise, we will learn about identifying data acquisition schedules, and preparing data by regularizing the schedule and interpolating missing locations. We will integrate these processes into a complete workflow where students will process original locations data into trajectories, and analyze them further in R.

EXERCISE: There and back again, part III: Extending PostGIS with Pl/R – Mathieu Basille

In this exercise, we will embed the R engine directly into the database to bring the power and features of R to PostGIS. In particular, we will learn the basics of Pl/R, before extending the home range concept in the database, and enabling complex treatments of trajectories, such as random steps along a trajectory, Brownian bridge kernels, or null models of movement.