# Report on the atRium Brno Training School

Margaux Depaermentier

## Overview

I have participated in the atRium Brno Training School (16th – 20th September 2024) at the Academy of Sciences at Brno. The workshop was particularly well organized both at the scientific and social level. Regarding the scientific part, the degree of difficulty sort of increased over the days, and every section was building on knowledge that was acquired in the previous days. It was therefore easy and satisfying to follow the logical development of the workshop and to be able to implement the new skills that were learned in the previous sections. Thematically, the sections were also very complementary and enabled to get insights into several aspects and possibilities of R applied to archaeological sciences. The first section provided us with an introduction to the basics of R, the second day introduced us to ARIADNE and SPARQL, the two following days went deeper into spatial analyses, and the last day of the workshop was dedicated to the implementation of all acquired knowledge using our own data (at least in my group, since another group – mostly made out of people who had no own data to work on – was deepening the topics related to ARIADNE and SPARQL). I will describe each section more thoroughly in the following paragraphs.

Before that, it is worth adding a few words about the social aspect of the workshop. Overall, the organization was extremely well mastered. Important information was clear and given on time, the team was incredibly kind and helpful, the breaks and social events were just amazing, and the organizers even helped with finding nice places for lunch or for the free time. This also worked thanks to the well-managed WhatsApp group of the summer school. This overall created a very friendly and welcoming atmosphere, which was of course very beneficial for the scientific part. The workshop being very intense and to some extent difficult, the general wonderful atmosphere of the workshop made it still easy to keep concentrated and motivated. I can add on top of this that the organizers also encouraged the participants to help each other during the workshop, and I have never seen this worked so good in any other academic context. Last but not least, in the context of heavy rain and flooding events, the organizing committee has even made great effort to help every participant finding train connections and their way to Brno, which is really remarkable.

## Day 1 – An introduction to R

The first day was dedicated to an introduction in R and enabled to make sure that all participants had at least a basic knowledge of RStudio and of the main commands that we would need in the following days. Petr Tkáč explained how the basic language of R worked and introduced us step by step with the essential definitions of vectors, values, objects, dataframes, and packages, using interactive examples. This introduction was very didactic and interactive. He then presented several functions that we would need in the following days, and while presenting, he gave us little tasks and exercises to make sure that we all understood what he presented and that we were all able to use the relevant codes.

A good aspect of this introduction was that in parallel to Petr Tkáč presentation, the other organizers and teachers of the workshop were helping the participants who had issues with the current tasks. The organisers introduced a very clever systems of stickers with the color-code "green" = "all good / done", "yellow" = "I have an approximative idea of what I am doing and work is in progress", and "red" = "I am lost, I need help", that we could stick on the top of our laptops. Hence, both the presenter, the helpers, and the participants had an overview of the overall progress in the group and it was possible to move on without losing anybody (either for being enable to solve the tasks or for being more advanced and hence quicker with the tasks).

In the introduction part, we therefore learned how to use R, how to create R projects and related folders, how to read data, how to explore and define them, how to make very preliminary statistics to describe the data and make first steps of analyses, and how to create our first plots. We received a lot of information about further literature and websites on which we can deepen the knowledge acquired in this introduction. We were provided with a dataset (e.g., representing the most basic data of all participants) to be able to test the functions that we were learning and to complete the different tasks and exercises.

The welcome reception at the end of the first day was very nice and contributed to create a great atmosphere between each participant of the workshop.

## Day 2 – ARIADNE AND SPARQL

The second day was presented by Petr Pajdla and split into three main parts: an introduction to ARIADNE, an introduction to SPARQL, and a whole afternoon dedicated to the practice. In the first part, we discovered the ARIADNE project and its interface. We learned how to best search for items and specify research areas and/or chronologies. In this context, we learned about the very useful "Getty AAT Subject", which provides a universal and multi-lingual definition of one specific item, object, person, place, or concept. Using Getty AAT Subject for a query or search enables to actually find all possible entry related to this object/item/person (etc.) regardless of the language in which the info is originally stored. Moreover, this is also very helpful for the following steps of this second workshop day, since using the identifier or the subject/object from the Getty AAT Subject makes sure that the meant object/subject will be used and that everybody can understand what is meant. Using the AAT hierarchies also helps finding internationally accepted translations for the needed object. Back to ARIADNE more specifically, we also learned how to really understand the results we get from the research and how to play with it. We also got sensitized to the fundamental aspect that each entry in ARIADNE is related to information about metadata, original publication, license, etc.

Since the ARIADNE platform offers both a chronological and a geographical dimension to the searched items, two further tools were presented to go deeper in the definition of periods and geographical areas. The PeriodO website was thus presented as an important complementary tool to define the chronological time frame of the period in absolute date, but also to depict the geographical distribution of the period as defined in a given area. In turn, the period identifier can be used in publications and research to make sure that the definition is clear for the whole community – and for the computers. The online tool "Geonames" may be considered the counterpart of PeriodO in term of geographical locations and their

nomenclature. Overall, ARIADNE represents a powerful tool to get information about existing data all around the world (although with a major focus on Europe) and over all existing periods.

As introduced in the second part of this second morning, this first level of information builds the fundament for Linked Open Data (LOD) in the Semantic Web, as they are used in the Uniform Resource Identifiers (URIs) to name and identify individual objects/subjects. This is in turn used to create so-called triples (in the form of "subject -> (predicate/property) -> object") that are used to create queries in, e.g. the Resource Description Framework (RDF) or SPARQL Protocol and RDF Query Language (SPARQL). In this context, we learned how to create prefix and abbreviations for each element of the triples using for example dbpedia and geonames for subjects and objects and owl or rdf-syntax for predicates/properties. Petr Pajdla showed us several examples of how to write triples, going step by step from the level using full URIs to the level using abbreviations only, to the level using turtle (Terse RDF Triple Language), in which triples have several levels of predicated/properties. He also explained how to integrate literals (i.e., values such as numbers, or dates) into the triples.

After the lunch break and the guided tour in the facilities of the Academy of Sciences (including library and archives), we used the afternoon to start applying this newly acquired knowledge, querying ARIADNE and SPARQL Endpoints. We discovered different types of queries and the various parts of the queries (prefix, selected variables, graph from which we are querying, endpoint of the query, limit and variables). We then spent the afternoon going step by step though a series of examples, with increasing difficulty and for which we were due to be more and more independent in the formulation of the queries. This was quite a challenge and I can imagine that the group of people who went deeper into these aspects on the last day got more opportunities to test their new skills and to try to integrate it for their own research interest.

## **Day 3 – Spatial analyses, part 1**

The third day was presented by Giacomo Bilotti. the third morning, he introduced us to spatial analyses in R. He presented the type of files that could be used in R (e.g., csv, GPKG, TIFF, ASCII), the packages that we would use in this workshop (mainly sf and terra, later also dplyr, tidyr, spData), the difference between raster and vectors (and the layers within the vectors) and their potential interactions, key aspects related to coordinate systems, as well as some of the main commands and basic operations that we can do with such data (e.g. subset, group, sum, summarise, create new columns, assign values, deal with NA values, etc.). We also learned how to create spatial objects and discovered few functions to, for examples, extract the extent of an object (using st_bbox), find data related to the research area (boundaries of the countries, DEM, etc.), calculate distances between points (using st_distance), buffer objects (using st_buffer), clip rasters, etc.

We further learned how to do operations with rasters, how to do basic stats (histograms, boxplots, etc.), to aggregate/disaggregate rasters, reclassify values, to crop rasters, etc. The most complicated part was to understand how and when to change the data into different types of objects to be used in various ways with the various packages. And I will need more time to play again and again with the provided scripts to get that point. Eventually, Giacomo Bilotti showed us various examples to do simple vizualisation and cartography. We created several maps and played with different tools to customize the maps and add more or less information and layers. We also learned how to create interactive maps. The most interactive

part of this third day was the moment when we went ourselves through the provided script and tried to fulfill various tasks and exercises. This was to some extent quite difficult and in this case, it would have been welcome to have more individual support (like in the first day) to be able to follow, understand, and manage everything.

On the evening of the fourth day, Michael Kempf and I gave a keynote lecture to present one case study in which we can integrate spatial analyses in bioarchaeological research. I first presented the background archaeological information, the research questions, the results of the isotope analyses carried out in this study and the reason why we need to integrate multivariate environmental analyses to go deeper in this case study. Michael Kempf presented the theoretical background, proxies, approaches, and methods of the environmental analyses, the different steps related to spatial analyses in R, as well as the results of this analysis. I finally developed on how to interpret the bioarchaeological data with respect to the model output presented by Michael Kempf.

## Day 4 – Spatial analyses, part 2

The fourth days was presented by Michael Kempf and dedicated to point pattern analyses, i.e., the relationship between point distribution (as vector) and other variables (usually as raster). After a short introduction on empirical versus theoretical models, we learned how to assess clustered, random or regular patterns using a set of comparison data and statistical tests (including Complete Special Randomness (CSR) with the Ripley's K-funtion and distribution test with KS-test, etc.

We could therefore estimate whether an observed point distribution (spatially) deviated from complete randomness at a given distance and then draw conclusions about why they do. To do so, we first needed to look at the second order intensity (or property) and to look if there are clusters, i.e., how sites are distributed in space, which is the point pattern itself. This implies to determine the extent of the window, i.e., the radius of the area of operation, which is sigma of the bandwidth. Which led to interesting discussions among participants. Even though we figured out that one can use statistical tests to determine this as well. Here, we observed the Kinhom clustered point distribution using Ripley's K (with respect to the Poisson process). We also used the KDE as a statistical technique to generate a smooth continuous distribution between data points that represent the density of the underlying pattern.

After defining the clusters, we worked on the first order of property, which introduces the covariates (such as topography, environmental settings, cultural background, etc.) or allows to compare periods with each other for example. We used the extract function in R to get first insights into the relationship between sites/objects and the different values of the background covariates. We used the rhohat function (from spatstat package) to get an idea of how the site/object are distributed compared to the surrounding and to measure the site intensity as a function of the underlying covariate. We also get sensitized to the fact that we should first understand the distribution of the covariate within the research area before trying to interpret the distribution of the sites in this area with respect to these covariates. The whole day was quite interactive, but the second part of the day was organized in a way that we had even more opportunities to play with the provided script and understand how it works step by step, by fulfilling tasks and resolving exercises.

## Day 5 – Application

As a first step of this fifth day, we discussed the best practice in data and code management and in how to re-use and cite the codes provided in the framework of this workshop. Then, after four very intensive days of theoretical and interactive teaching, we came to the point where we could apply our new skills in our own research. The participants were divided into two groups and since I had intentionally brought my own data to Brno, I integrated the group that was due to do spatial analyses with their own data. This was a very welcome opportunity to start working on real data and to get help at the moment when we get confronted with the first technical issues. The first tremendous task for me consisted in finding background data for my huge research area (Northern, Eastern, Central, and Southern Europe) and to make them adequate for use with the scripts we received in the workshop. This clearly showed me that some steps that seemed to be clear and easy in theory were considerably more difficult in praxis, because they required more background knowledge about R, R-packages, and data format. I tried to re-use both Michael's and Giacomo's scripts with my own data, but got stuck several times due to my lack of knowledge in this context. It was a luck that the organizers were still available to help us. But the group size became at this very moment an issue, since too many diverse problems occurred and everybody basically constantly needed individual help. I finally realized that the data I prepared was not perfectly adequate for the function that we learned in the workshop, and that this was a reason why some parts of the codes were not applicable to my dataset. By the time I was done with preparing my dataset, the workshop was over, but I learned a lot and I feel prepared to continue working on it with my own data. With the knowledge acquired in this summer school, I also feel more confident about the fact that I may be able to re-use codes provided in publications of interest for my research.

## Closing remarks

I would therefore like to thank the organizers of the atRium Brno Training School to have created this unique opportunity to learn about basics in R and especially in spatial analyses in R, and to have introduced us to ARIADNE and SPARQL. I would never have managed to get so much knowledge and skills on these topics within a single week if I had needed to read about it on my own. I am currently working in several bioarchaeological projects (including isotope data) in which it is crucial to compared the spatial distribution of our observed data (related to dietary and subsistence practices) to underlying environmental and cultural covariates such as information on slope, elevation, longitude and latitude, wetness or aridity, and further environmental settings as well as the distribution area of cultural groups and those of specific practices. I now know about several approaches to investigate these aspects and I look forward to apply these new skills in my research. This workshop filled an important gap in my academic profile and I am extremely grateful to have been selected to participate in this summer school.

Dr. Margaux L. C. Depaermentier
Postdoctoral researcher
Vilnius University, Lithuania