$$\int_{2012}^{\infty} ma\,\mathbf{ptm}$$

K. Bartoszek (iD) * (Linköping)

Revisiting the Nowosiółka skull with **RMaCzek**

Abstract One of the first fully quantitative distance matrix visualization methods was proposed by Jan Czekanowski at the beginning of the previous century. Recently, a software package, **RMaCzek**, was made available that allows for producing such diagrams in R. Here we reanalyze the original data that Czekanowski used for introducing his method, and in the accompanying code show how the user can specify their own custom distance functions in the package.

2010 Mathematics Subject Classification: Primary: 62H99; Secondary: 62-04; 92B10..

Key words and phrases: Czekanowski's diagram, craniometry, human evolution, multivariate distance methods..

1. Introduction: Czekanowski's diagram Czekanowski's diagram is thought to be one of the first taxonomic and proximity visualization methods. It was proposed by the Polish anthropologist and statistician Jan Czekanowski in 1909 [4]. In order to construct such a diagram one needs to be able to calculate the distance between each pair of observations. Czekanowski used the average difference between the attributes of two d–dimensional objects, $\vec{x}$ and $\vec{y}$,

$$DD(\vec{x}, \vec{y}) = \frac{1}{d}\sum_{r=1}^{d}|\vec{x}_r - \vec{y}_r|. \tag{1}$$

Then, one needs to solve a seriation problem [7]—find an arrangement of the observations such that objects close together under the distance are close together when placed on a straight line. Afterwords, one can represent the data as a matrix, where each cell is to represent the calculated distance between the row and column object. Today heatmaps achieve this through a direct colour gradient, but in Czekanowski's times monochrome graphics were substantially easier to produce on a large scale (but, e.g., in 1873 a summary

---

matrix of 40 separate maps of Paris showing various characteristics was presented in colour [9, 17]). Hence, each cell is made up of a symbol representing the distance. The number of possible symbols is limited—hence the distances are grouped. The distance's range is divided into consecutive subintervals, with a unique symbol assigned to each. For example, this can be a black dot with varying size—the smaller the distance the bigger the dot—see e.g. Fig. 1. Today, Czekanowski's method is classified under seriation, matrix reordering and visualization methods. Closely, related to this are the matrix re–ordering and visualization methods proposed by Bertin (e.g. [3]). The main difference is that Czekanowski's approach works a similarity matrix, with rows and columns columns corresponding to the same objects (with the same permutation of them), while Bertin's focuses on presenting a matrix (e.g. containing the actual measurements for the different objects) with possibly separate rearrangements of the rows and columns. In today's world, colour heatmaps and clustering methods are a common choice for visualizing matrices (be it for those containing similarity scores or actual measurements). This is also due to their rapid development, alongside the increasing availability of of more and more sophisticated computer graphics. Currently a number of R [10] heatmap packages are available on CRAN or Bioconductor (e.g., `stats::heatmap()`, **ComplexHeatmap** [5, 6] **heatmap3** [18], **heatmaps**, **pheatmap**, to name a few). For a thorough review, and history of this field we refer the reader to, e.g., [8, 17].

One can immediately notice that the creation of Czekanowski's diagram does not depend on the actual way the distance was calculated—any function from the Cartesian product of the observation's space with itself to the set of non–negative real numbers (or even more generally to the space of symbols representing distances) would suffice. Given Czekanowski's particular dataset (craniometric data of archaic humans) the distance of Eq. (1) sufficed. However, if one would have categorical observations, would want to take into account correlations between attributes, weights of attributes or focus on alternative aspects (as we will in Section 3), then other distance functions could be more appropriate.

There are a number of previous software implementations of Czekanowski's diagram. The most well known one was the Visual Basic **MaCzek** program [13]. Then, following encouragement from Mirosław Krzyśko at the XXIV National Conference Applications of Mathematics in Biology and Medicine in Zakopane–Kościelisko, 2018 and afterwords from Arkadiusz Sołtysiak, Albin Västerlund [16] implemented the R package **RMaCzek**. The package allows for the creation of Czekanowski's diagram under user provided distance functions and seriation methods. From our *in–silico* experiments [2, 16] it turned out that the most effective seriation method is based on finding a Hamiltonian path in the full graph of the observations (with edge lengths equalling the distances between the observations) that is minimal [1].

In this work we reanalyze the original data that introduced Czekanowski's diagram to the world. We also provide scripts[1] that show how to use **RMaCzek**[2]. A key part of the scripts are the examples that show how to provide a custom user defined distance function with arbitrary control parameters. It is important to point out that comparing populations through cranial measurements was dropped in the 1960s thanks to the advancement in genetics that showed that many morphological traits are environmentally controlled and not genetically [14]. We underline, that here we do not aim at drawing conclusions concerning archaic human populations but at replicating analyses from over a century ago.

2. The Nowosiółka skull At the beginning of the previous century the debate on how humans developed to today was fuelled by numerous fossil finds. Two hypotheses can be seen competing [15]. The first one, ascribed to Schwalbe [15], in modern language, stated that *Homo neanderthalensis* went completely extinct by the Paleolithic, not leaving behind any intermediate forms to *Homo sapiens*. The other, favoured by Stołyhwo [15] was that neanderthalic features survived the Paleolithic and could also be present during the Era of History. He decided to disprove Schwalbe through a semi–quantitative approach by comparing measurements of a Scythian, ca 30–year old warrior's skull [15] found in Nowosiółka (obviously *H. sapiens*) with those of neanderthalic ([15] after Schwalbe) skulls (those found in the caves of Neandertal, Spy and Krapina). Czekanowski [4] performed a fully quantitative analysis of the craniometric data and found the Nowosiółka skull to be placed in the *H. sapiens*' cluster.

3. Reanalysis with RMaCzek With the availability of the **RMaCzek** package we will attempt to replicate both Stołyhwo's and Czekanowski's studies, do further data exploration and see whether any of two hypotheses concerning the survival of neanderthalic features is better supported. Stołyhwo presented measurements of 47 cranial features, some of them multivariate, from 31 human skulls. While this might seem a lot, most of the measurements are missing, and most skulls have only a few features measured. For some skulls a single value is provided for a given feature, while for others the range of the feature is provided. Finally, in some cases it is not possible to assign a feature to any particular skull, it is just written "piece of skull from Krapina". For the Nowosiółka skull all features are measured. Then, the Nowosiółka skull's measurements are compared with neanderthalic (called *Homo primigenius* in [15], today this nomenclature is deprecated and *H. neanderthalensis* is used instead) ones. Each feature is classified either as not–different, similar to or different from *H. primigenius*'s respective feature. Unfortunately it is not stated explicitly what computational procedure is employed for this

---

[1] https://github.com/krzbar/RMaCzek_KKZMBM2020 (with craniometric measurements)
[2] https://cran.r-project.org/web/packages/RMaCzek/

classification. Czekanowski, employed the approach described in Section 1, with the distance of Eq. (1), to a subset of 13 skulls, using 27 of the features. He cites [15] as the data source. The choice of skulls and features is not explained in [4], but it seems that the skulls with more measurements were chosen and the features are such that they are present in the Neandertal, Brüx (*H. sapiens* representative) and Nowosiółka skulls. Czekanowski does not describe the rationale for his arrangement of the skulls but it could be guided by the timeline presented in [12], on p. 14.

We first digitalized all the data presented in [15] and they are made available alongside the R scripts for this work. For further analyses we keep to the 13 skulls of [4]. In these 13 skulls 64.6% of measures variables considered by Stołyhwo are missing, if we restrict ourselves to the 27 variables considered in [4] we have "only" 29.9% missingness. However there are still substantial missing value levels amongst the individual skulls: Spy I (7.4%), Spy II (14.8%), Krapina C (40.7%), Krapina D (59.3%), Neandertal (0%), Gibraltar (48.1%), Pithecanthropus (22.2%), Kannstatt (18.6%), Galey Hill (55.6%), Brunn (55.6%), Brüx (0%), Egisheim (66.7%) and Nowosiółka (0%). It was noticed [13] that observations that have missing values on more than half the variables should not be used, as they may be close to multiple, often very different from each other observations. Hence in our data set missingness is a potential serious problem.

Then, our first goal was to replicate the distance matrix (Tabelle II in [4]) and diagram (Tabelle III in [4]) between in the skulls that can be found in [4]. We illustrate this in Fig. 1. We can see that our derived distance matrix differs a bit from the one Czekanowski presented in [4]. However, all but a few cells differ by less than 4%. Certainly a substantial amount of the difference can be attributed to rounding errors, Czekanowski had to commit them when calculating manually or with a mechanical calculator, and also his distances are presented up to the third decimal point. However, it is difficult to discover what the larger discrepancies are due to. We do not know if he only used the measurements presented in [15], or maybe some additional ones. Furthermore, we also do not know how variables with only an upper and lower bound provided were treated. Here we take the arithmetic average of the maximum and minimum of the range. It is worth pointing out that in Czekanowski's original distance matrix there is an obvious typo in the distance between the Neandertal and Galley Hill skulls. In Fig. 1 we plot the resulting Czekanowski's diagram for the best found arrangement of the skulls by **RMaCzek**. In [2] we provided the between–skulls distance matrix from [4] to **RMaCzek** and obtained a better (under all objective functions) permutation than in [4], however all the qualitative conclusions were the same. Here, we find the same arrangement, derived directly from the raw measurements and we obtain the same figure as in [2].

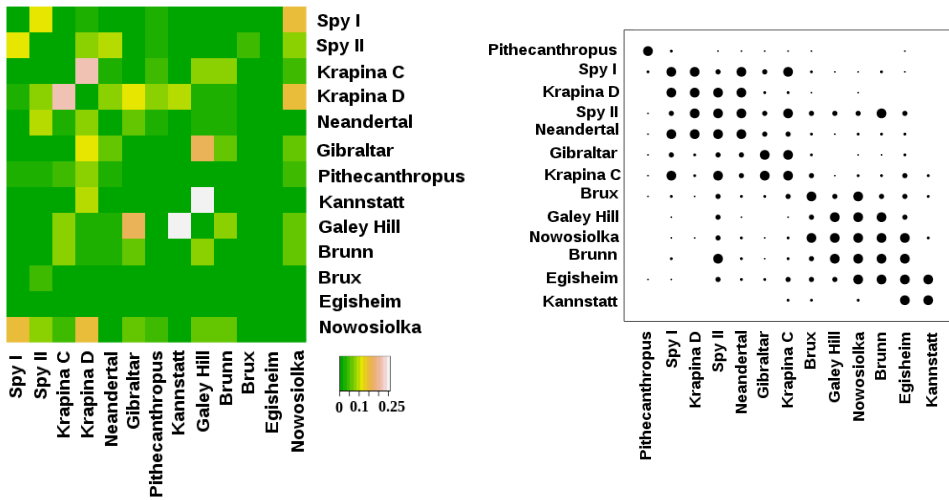We now turn to finding the placement of the Nowosiółka skull with respect

Figure 1: Left: absolute relative difference between skulls' distance matrix found in [4] and the distance matrix calculated using **RMaCzek**, right: Czekanowski's diagram found using **RMaCzek**.

to the other skulls. We consider three distances—the DD distance of Eq. (1), the squared Euclidean distance and counting the number of variables supporting different classification (*H. neanderthalensis* or *H. sapiens*). The first and the last one are those considered by Czekanowski and Stołyhwo, respectively, and hence relevant for our attempt to replicate the results. We then take the Euclidean distance function for comparison. We also try out a number of other options. We keep angles measured in degrees and also convert them to radians. Some of the variables are ratios of other variables—they are dependent and this could interfere with the seriation procedure. Hence, apart from all the variables, we also perform analyses with the ratios removed and alternatively where the ratio variables are kept, but their components are removed. We also take into consideration all the variables, or only the 27 considered in [4]. We take all observations and also removed those with more than 50% missingness. We also normalize, mean centre and divide by standard deviation, all the variables.

Furthermore, we wanted to see if we can replicate the conclusion of [15], that Nowosiółka is more similar to neanderthalic skulls. To do this we needed to implement a distance function that could mimic the table on p. 25 in [15]. Stołyhwo considered whether each variable measured in the Nowosiółka skull was the same, similar or different from *H. primigenius* ones. To mimic this we employ the following procedure. We take two focal skulls—Neandertal (representing the *H. neanderthalensis* clade) and Brüx (representing the *H. sapiens* clade). We take this pair as they have measurements on all 27 variables considered in [4]. Then, to calculate the distance between observations x and

y, with `v1` and `v2` being the measurements of the representatives of the two clades, the below R procedure is used.

```
x1<-abs(x-v1);x2<-abs(x-v2);y1<-abs(y-v1);y2<-abs(y-v2)
v_x<-x1-x2;v_y<-y1-y2
qx<-quantile(v_x,probs=c(1/3,2/3),na.rm=TRUE)
qy<-quantile(v_x,probs=c(1/3,2/3),na.rm=TRUE)
xd<-rep(NA,length(v_x));yd<-rep(NA,length(v_y))
xd[which(v_x<=qx[1])]<-0;xd[which(v_x>qx[2])]<-2
xd[intersect(which(v_x>qx[1]),which(v_x<=qx[2]))]<-1
yd[which(v_y<=qy[1])]<-0;yd[which(v_y>qy[2])]<-2
yd[intersect(which(v_y>qy[1]),which(v_y<=qy[2]))]<-1
distxy<-mean(abs(xd-yd),na.rm=TRUE)
```

We may recognize that what is done is that for both `x` and `y` each variable is classified as being closer to the respective variable in `v1` or `v2` (or in–between). Then, we take the average of how these patterns differs between `x` and `y`. Importantly, under this distance normalization of the variables is not done.

All together this resulted in 96 possible setups (with resulting arrangements of the skulls). We went through this list manually and in 65 setups the Nowosiółka skull was placed closer to the *H. neanderthalensis* skulls than in the original result presented in Fig. 1. These cases can be divided as those where the Nowosiółka skull was: placed on the boundary between *H. sapiens* and *H. neanderthalensis* skulls (11 setups, exemplary setup number 13 in Fig. 2); separated from the *H. sapiens* skulls by the Kannstatt skull (which in the original presentation can be thought to be a singleton) and followed by *H. neanderthalensis* skulls (6 setups, exemplary setup number 21 in Fig. 2); on the border with *H. neanderthalensis* skulls of a partial *H. sapiens* skulls set (2 setups, exemplary setup number 82 in Fig. 3); on the border of the whole *H. sapiens* skulls set which was placed inside the *H. neanderthalensis* skulls (4 setups, exemplary setup number 66 in Fig. 3); jointly with the Kannstatt and Brüx skulls placed amongst *H. neanderthalensis* skulls (23 setups, exemplary setup number 25 in Fig. 2); jointly with the Brüx skull placed amongst *H. neanderthalensis* skulls (1 setup, number 29 in Fig. 3); a singleton followed by *H. neanderthalensis* skulls (18 setups, exemplary setups number 26 and 92 in Figs. 2 and 3). Unrelated to this, but interestingly, the Brüx skull was placed in two setups as a singleton inside the *H. neanderthalensis* skulls (number 46 in Fig. 3). No particular pattern was observed in the setups related to the above, in particular there seems to be no dependence on the distance function (20 are Stołyhwo's, 23 L^2 and 22 DD distances). However, some observations from the exemplary graphs in Figs. 2 and 3, are presented in the Discussion. All the results are published alongside the source code. We also tested, what the results would be on the raw data, i.e. without normalization. The Nowosiółka skull would be placed more firmly within the *H. sapiens* skulls—in particular no singleton followed by *H. neanderthalensis* skulls was observed.
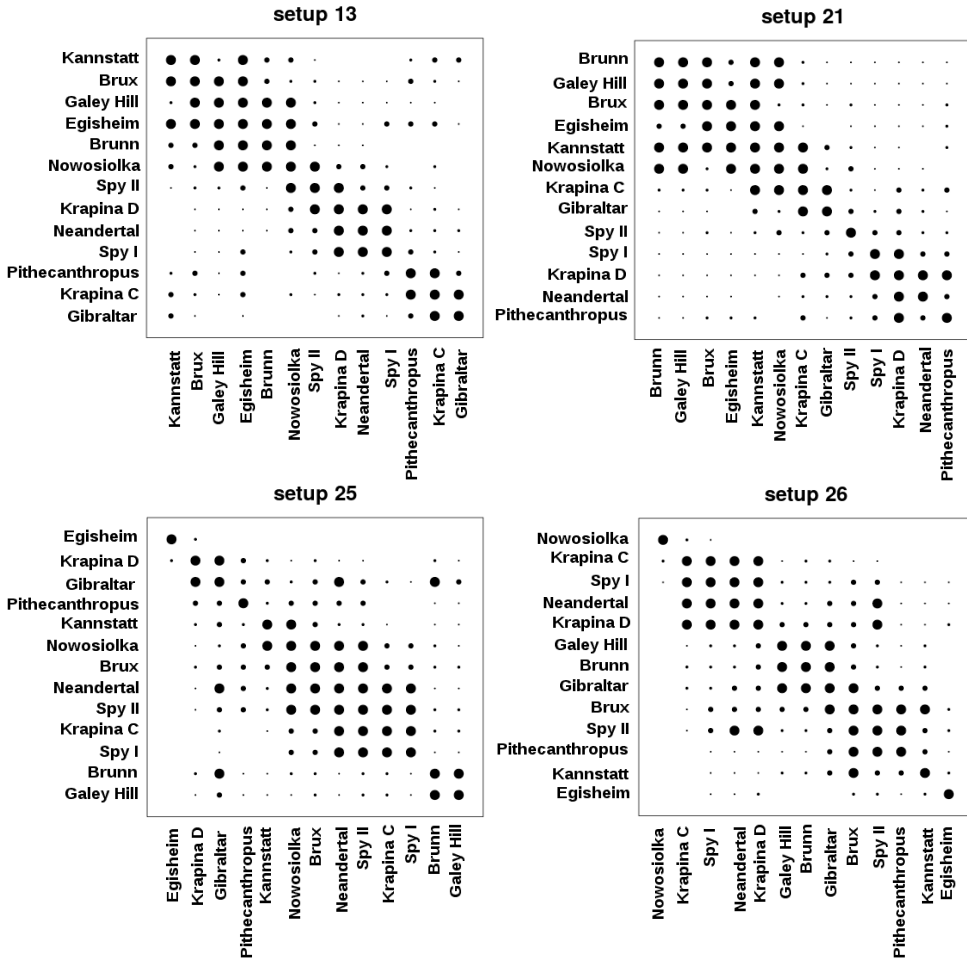
Figure 2: Exemplary (part 1) Czekanowski's diagrams, where the Nowosiółka skull was placed closer to the *H. neanderthalensis* skulls than in the original analysis by Czekanowski [4] and **RMaCzek** with Czekanowski's settings.
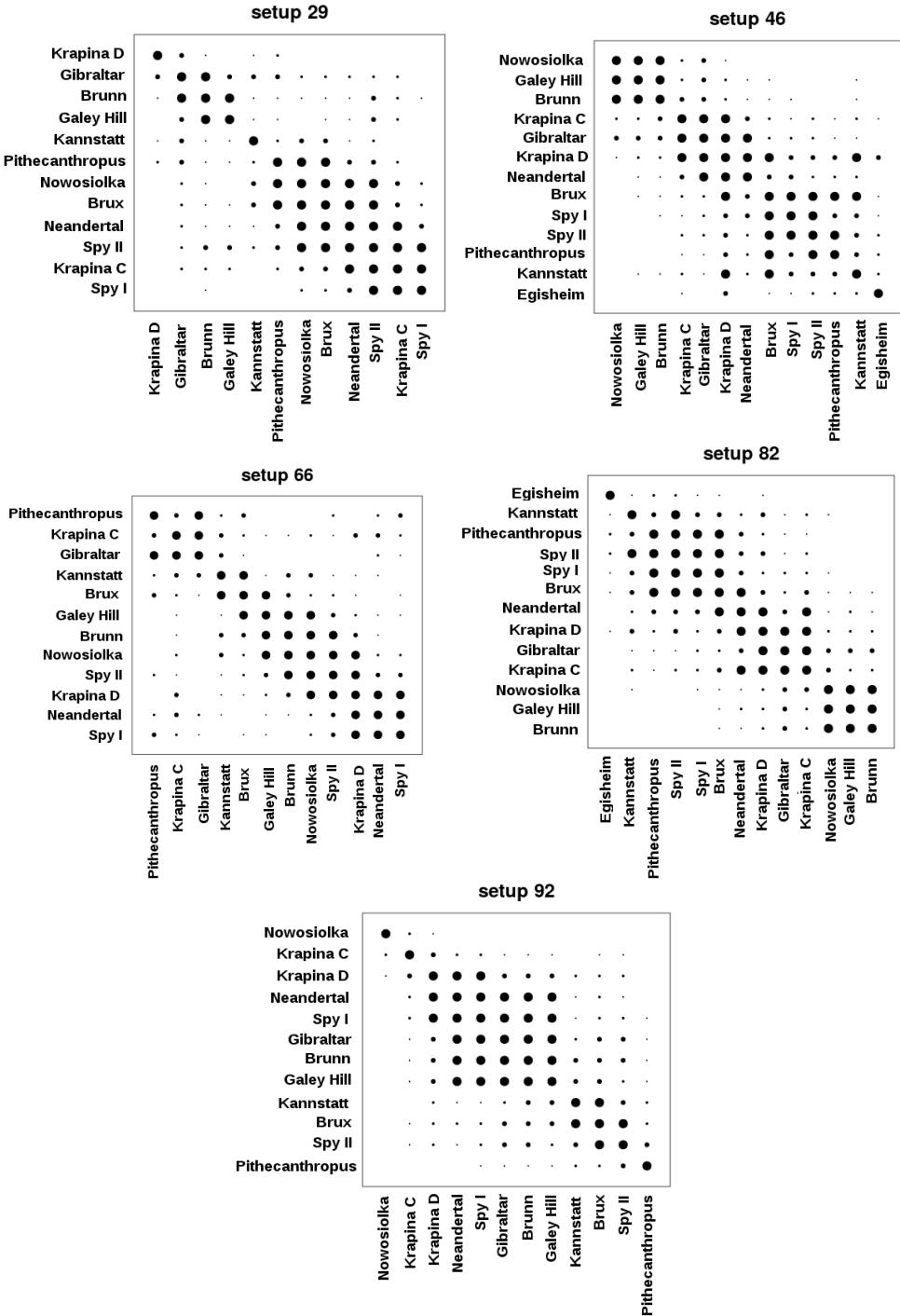
Figure 3: Exemplary (part 2) Czekanowski's diagrams, where the Nowosiółka skull was placed closer to the *H. neanderthalensis* skulls than in the original analysis by Czekanowski [4] and **RMaCzek** with Czekanowski's settings.

4. Discussion The work here has two aims. To show what possibilities **RMaCzek** has when creating Czekanowski's diagrams and also to return to the original data that was used to introduce the method to the world. Since Czekanowski used the method to challenge Stołyhwo's conclusions concerning the evolutionary relatedness of the Nowosiółka skull we wanted to see which claim would modern methods substantiate and also what approaches could support the alternative claim.

Using various distance functions, variables and other settings it was noticed that the placement of the Nowosiółka skull can greatly vary. However, in only 18 out of the 96 considered settings it can be thought to be classified as *H. neanderthalensis* instead of *H. sapiens*, in line with Stołyhwo's hypothesis. Interestingly none of these results, where the Nowosiółka skull is closest to the *H. neanderthalensis*, were under Stołyhwo's distance function. Even in the cases (in Figs. 2 and 3) where the Nowosiółka skull was found closer to the *H. neanderthalensis* skulls it was not grouped with them. In setups 26 and 92 the Nowosiółka skull is an outlier. In the other setups it is either just placed as bordering skull, displaying weak similarity to the *H. neanderthalensis* skulls (setup 82), or displaying similar similarity to *H. sapiens* and *H. neanderthalensis* skulls (setups 13, 21, 66). Alternatively, it was grouped with a *H. sapiens* skull in the midst of *H. neanderthalensis* skulls (setup 29).

From the multitude of obtained permutations, one must conclude that with so few observations and high level of missing values the problem is very sensitive. It is of course impossible to completely replicate the analyses from over 110 years ago. In particular we do not know how Stołyhwo classified each variable with respect to what is said about the Nowosiółka skull's placement. Perhaps with a different choice of skulls to be the "standard" we would have obtained a different result.

It must be pointed out that in this work we have restrained ourselves to data used by both Czekanowski and Stołyhwo. It is possible that they could have also tacitly supported their works with data from other contemporary to them sources, e.g. [12]. Given today's technology one should compare the extractable DNA from the Nowosiółka skull with those of Neanderthals' and *H. sapiens'*. This would definitely solve the problem. We are not aware if this particular skull has been sequenced, but certainly this would be beyond the scope of this work, even if the actual specimen is still available somewhere.

While we do not have as our aim to draw conclusions concerning archaic human populations it is worth commenting that our analyses do not contradict Czekanowski. The, Scythian, Nowosiółka skull, seems to be related to the *H. sapiens* skulls under most settings. On the other hand, Stołyhwo's hypothesis was that neanderthalic features did survive into the Era of History, perhaps strongly deformed. What we know today, is that neanderthalic DNA is present in non–sub–Saharan modern human populations [11], proving, af-

ter a whole century, Stołyhwo correct, albeit in a way he could not had have foreseen at the time.

5. Software Availability RMaCzek can be found at `https://cran.r-project.org/web/packages/RMaCzek/` and `https://github.com/krzbar/RMaCzek/`. The R scripts allowing for the replication of the work here and the craniometric measurements from [15] are available at `https://github.com/krzbar/RMaCzek_KKZMBM2020`. The above scripts show how to code user defined distance functions for further usage by **RMaCzek**.

6. Acknowledgments KB is grateful to Arkadiusz Sołtysiak for valuable comments.

7. References

[1] Z. Bar-Joseph, E. D. Demaine, D. K. Gifford, and T. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(1):22–29, 2001. URL `https://doi.org/10.1093/bioinformatics/17.suppl_1.S22`. Cited on p. 256.

[2] K. Bartoszek and A. Västerlund. "Old Techniques for New Times": the **RMaCzek** package. *Biometrical Letters*, 57(2):89–118, 2020. URL `https://doi.org/10.2478/bile-2020-0008`. Cited on pp. 256 and 258.

[3] J. Bertin. *Sémiologie Graphique*. Gauthier–Villars, Mouton, Paris, 1967. Cited on p. 256.

[4] J. Czekanowski. Zur Differentialdiagnose der Neandertalgruppe. *Korespondentblatt der Deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte*, XL(6/7):44–47, 1909. Cited on pp. 255, 257, 258, 259, 261, and 262.

[5] Z. Gu. Complex heatmap visualization. *iMeta*, 1:e43, 2022. URL `https://doi.org/10.1002/imt2.43`. Cited on p. 256.

[6] Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016. URL `https://doi.org/10.1093/bioinformatics/btw313`. Cited on p. 256.

[7] M. Hahsler, K. Hornik, and C. Buchta. Getting things in order: An introduction to the R package **seriation**. *J. Stat. Softw.*, 25(3):1–34, 2008. URL `https://doi.org/10.18637/jss.v025.i03`. Cited on p. 255.

[8] I. Liiv. Seriation and matrix reordering methods: An historical overview. *Stat. Anal. Data. Min.*, 3(2):70–91, 2010. URL https://doi.org/10.1002/sam.10071. Cited on p. 256.

[9] T. Loua. *Atlas statistique de la population de Paris.* J. Dejey, Paris, 1873. Cited on p. 256.

[10] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/. Cited on p. 256.

[11] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich. The genomic landscape of neanderthal ancestry in present–day humans. *Nature*, 507(7492):354–357, 2014. URL https://doi.org/10.1038/nature12961. Cited on p. 263.

[12] G. A. Schwalbe. *Studien zur Vorgeschichte des Menschen.* E. Nägele, Stuttgart, 1906. Cited on pp. 258 and 263.

[13] A. Sołtysiak and P. Jaskulski. Czekanowski's diagram. A method of multidimensional clustering. In *New Techniques for Old Times. CAA 98. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 26th Conference, Barcelona, March 1998*, pages 175–184. Oxford, number 757 in BAR International Series, 1999. URL https://proceedings.caaconference.org/paper/29_soltysiak_jaskulski_caa_1998/. Cited on pp. 256 and 258.

[14] A. Sołtysiak and T. Kozłowski. Komentarz do identyfikacji *Cranium* 13/05 z Fromborka jako kości Mikołaja Kopernika. *Archeologia Polski*, LIV(2):281–290, 2009. URL http://antropologia.uw.edu.pl/AS/as-026.pdf. Cited on p. 257.

[15] K. Stołyhwo. Czaszka z Nowosiółki jako dowód istnienia w okresie historycznym kształtów pokrewnych z *Homo primigenius. Rozprawy Wydziału matematyczno–przyrodniczego Akademii Umiejętności*, XLVIII(B): 1–27, 1908. Cited on pp. 257, 258, 259, and 264.

[16] A. Västerlund. Czekanowski's Diagram: Implementing and exploring Czekanowski's Diagram with different seriation methods, 2019. MSc thesis, Division for Statistics and Machine Learning, Department of Computer and Information Science, Linköping University. Cited on p. 256.

[17] L. Wilkinson and M. Friendly. The history of the cluster heat map. *Am. Stat.*, 63(2):179–184, 2009. URL https://doi.org/10.1198/tas.2009.0033. Cited on p. 256.

[18] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr. Advanced heat map and clustering analysis using heatmap3. *Adv. Comput. Genom.*, page Article ID 986048, 2014. URL https://doi.org/10.1155/2014/986048. Cited on p. 256.

Analiza czaszki z Nowosiółki z użyciem pakietu RMaCzek.
Krzysztof Bartoszek

Streszczenie Na początku zeszłego stulecia Jan Czekanowski, polski antropolog oraz statystyk, zaproponował jedną z pierwszych obiektywnych metod uporządkowania oraz zobrazowania macierzy odległości. W 2019 roku został opracowany oraz udostępniony pakiet **RMaCzek**, który pozwala na tworzenie diagramów Czekanowskiego w środowisku R. W niniejszej pracy dokonano ponownej analizy danych, które posłużyły Czekanowskiemu do zaprezentowania własnej metody oraz zaproponowano, jak w pakiecie **RMaCzek** użytkownik może wprowadzać własną funkcję odległości.

2010 Klasyfikacja tematyczna AMS (2010): 62H99; 62-04; 92B10.

Słowa kluczowe: diagram czekanowskiego, kraniometria, metody odległości wielocechowych, rozwój ludzkości.

Krzysztof Bartoszek[a] (ur. 1984 w Bydgoszczy), obecnie wykładowca statystyki na Uniwersytecie w Linköpingu. Absolwent informatyki Politechniki Gdańskiej (mgr inż.), biologii obliczeniowej Uniwersytetu w Cambridge (MPhil). Doktorat z statystyki matematycznej pod kierunkiem Serika Sagitova uzyskał w 2013 r. na Uniwersytecie w Göteborgu. Jego główne zainteresowania związane są z procesami stochastycznymi w filogenetyce.

[a]References to his research papers are found in MathSciNet under ID: 793554 and the European Mathematical Society, FIZ Karlsruhe, and the Heidelberg Academy of Sciences bibliography database known as zbMath under ai:Bartoszek.Krzysztof.

K. Bartoszek
Linköping University
Department of Computer and Information Science
The Division of Statistics and Machine Learning
Linköping University, 581 83, Linköping, Sweden
E-mail: krzysztof.bartoszek@liu.se; krzbar@protonmail.ch