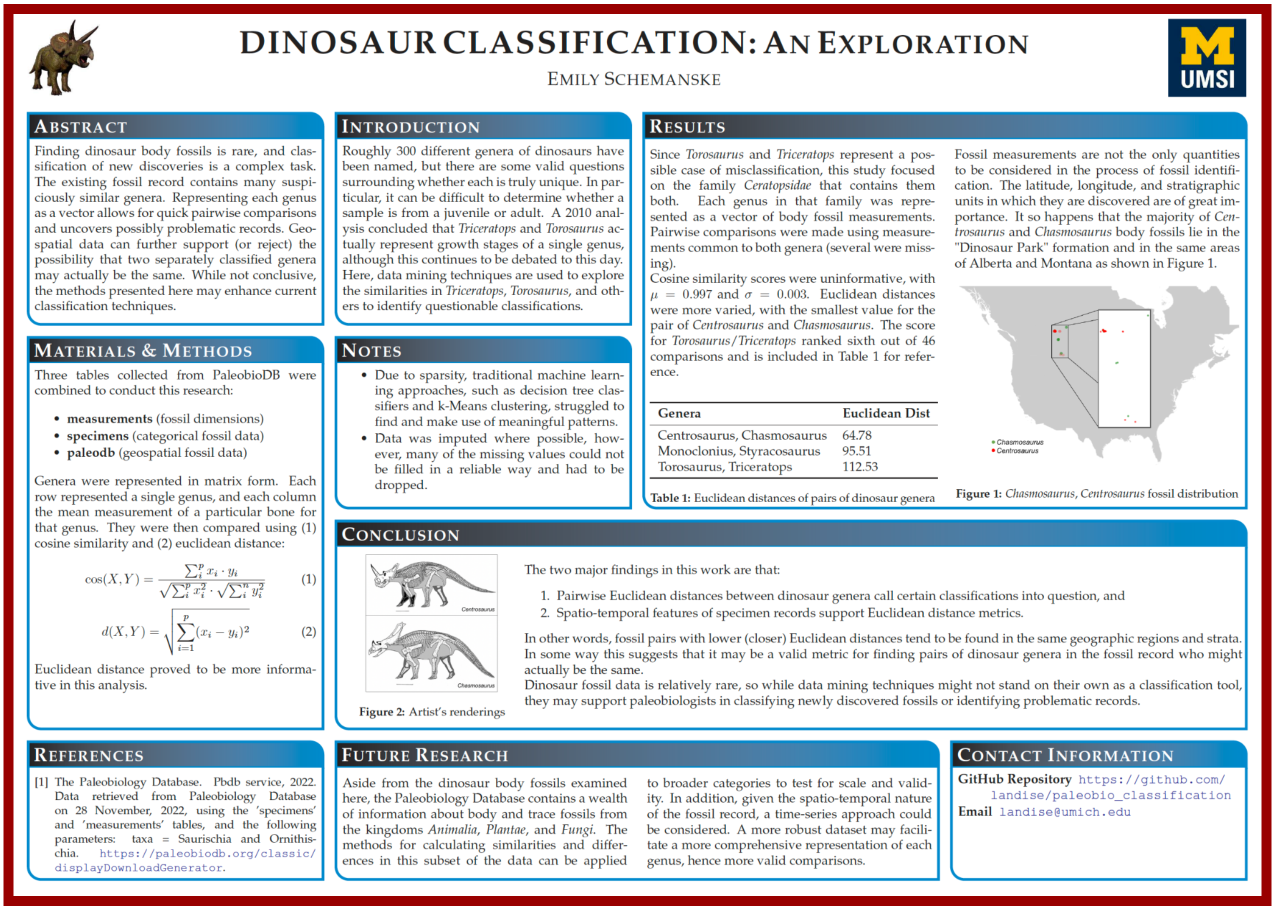


DINOSAUR CLASSIFICATION: AN EXPLORATION

EMILY SCHEMANSKE

UMSI

ABSTRACT

Finding dinosaur body fossils is rare, and classification of new discoveries is a complex task. The existing fossil record contains many suspiciously similar genera. Representing each genus as a vector allows for quick pairwise comparisons and uncovers possibly problematic records. Geospatial data can further support (or reject) the possibility that two separately classified genera may actually be the same. While not conclusive, the methods presented here may enhance current classification techniques.

MATERIALS & METHODS

Three tables collected from PaleobioDB were combined to conduct this research:

- **measurements** (fossil dimensions)
- **specimens** (categorical fossil data)
- **paleodb** (geospatial fossil data)

Genera were represented in matrix form. Each row represented a single genus, and each column the mean measurement of a particular bone for that genus. They were then compared using (1) cosine similarity and (2) euclidean distance:

$$\cos(X,Y) = \frac{\sum_i^p x_i \cdot y_i}{\sqrt{\sum_i^p x_i^2} \cdot \sqrt{\sum_i^n y_i^2}} \qquad (1)$$

$$d(X,Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \qquad (2)$$

Euclidean distance proved to be more informative in this analysis.

INTRODUCTION

Roughly 300 different genera of dinosaurs have been named, but there are some valid questions surrounding whether each is truly unique. In particular, it can be difficult to determine whether a sample is from a juvenile or adult. A 2010 analysis concluded that *Triceratops* and *Torosaurus* actually represent growth stages of a single genus, although this continues to be debated to this day. Here, data mining techniques are used to explore the similarities in *Triceratops*, *Torosaurus*, and others to identify questionable classifications.

NOTES

- Due to sparsity, traditional machine learning approaches, such as decision tree classifiers and k-Means clustering, struggled to find and make use of meaningful patterns.
- Data was imputed where possible, however, many of the missing values could not be filled in a reliable way and had to be dropped.

RESULTS

Since *Torosaurus* and *Triceratops* represent a possible case of misclassification, this study focused on the family *Ceratopsidae* that contains them both. Each genus in that family was represented as a vector of body fossil measurements. Pairwise comparisons were made using measurements common to both genera (several were missing).
Cosine similarity scores were uninformative, with $\mu = 0.997$ and $\sigma = 0.003$. Euclidean distances were more varied, with the smallest value for the pair of *Centrosaurus* and *Chasmosaurus*. The score for *Torosaurus*/*Triceratops* ranked sixth out of 46 comparisons and is included in Table 1 for reference.

Genera	Euclidean Dist
Centrosaurus, Chasmosaurus	64.78
Monoclonius, Styracosaurus	95.51
Torosaurus, Triceratops	112.53

Table 1: Euclidean distances of pairs of dinosaur genera

Fossil measurements are not the only quantities to be considered in the process of fossil identification. The latitude, longitude, and stratigraphic units in which they are discovered are of great importance. It so happens that the majority of *Centrosaurus* and *Chasmosaurus* body fossils lie in the "Dinosaur Park" formation and in the same areas of Alberta and Montana as shown in Figure 1.

Figure 1: *Chasmosaurus*, *Centrosaurus* fossil distribution

CONCLUSION

Figure 2: Artist's renderings

The two major findings in this work are that:

1. Pairwise Euclidean distances between dinosaur genera call certain classifications into question, and
2. Spatio-temporal features of specimen records support Euclidean distance metrics.

In other words, fossil pairs with lower (closer) Euclidean distances tend to be found in the same geographic regions and strata. In some way this suggests that it may be a valid metric for finding pairs of dinosaur genera in the fossil record who might actually be the same.
Dinosaur fossil data is relatively rare, so while data mining techniques might not stand on their own as a classification tool, they may support paleobiologists in classifying newly discovered fossils or identifying problematic records.

REFERENCES

[1] The Paleobiology Database. Pbdb service, 2022. Data retrieved from Paleobiology Database on 28 November, 2022, using the 'specimens' and 'measurements' tables, and the following parameters: taxa = Saurischia and Ornithischia. https://paleobiodb.org/classic/displayDownloadGenerator.

FUTURE RESEARCH

Aside from the dinosaur body fossils examined here, the Paleobiology Database contains a wealth of information about body and trace fossils from the kingdoms *Animalia*, *Plantae*, and *Fungi*. The methods for calculating similarities and differences in this subset of the data can be applied to broader categories to test for scale and validity. In addition, given the spatio-temporal nature of the fossil record, a time-series approach could be considered. A more robust dataset may facilitate a more comprehensive representation of each genus, hence more valid comparisons.

CONTACT INFORMATION

GitHub Repository https://github.com/landise/paleobio_classification
Email landise@umich.edu

Dinosaur Classification: An Exploration

Emily Schemanske

Abstract

Finding dinosaur body fossils is rare, and classification of new discoveries is a complex task. The existing fossil record contains many suspiciously similar genera. Representing each genus as a vector allows for quick pairwise comparisons and uncovers possibly problematic records. Geospatial data can further support (or reject) the possibility that two separately classified genera may actually be the same. While not conclusive, the methods presented here may enhance current classification techniques.

Introduction

Roughly 300 different genera of dinosaurs have been named, but there are some valid questions surrounding whether each is truly unique. In particular, it can be difficult to determine whether a sample is from a juvenile or adult. Inability to discriminate between the two can have “potentially enormous implications - juveniles and adults of the same taxon may be misidentified as adults of different species, affecting taxonomic and phylogenetic hypotheses (Hone et al.). Take, for example, the case of *Triceratops* and *Torosaurus*. In 2010, two paleontologists from Montana State University published a paper stating that “although they have been considered distinct genera for over a century, ontogenetic analyses reveal that *Triceratops* and *“Torosaurus”* actually represent growth stages of a single genus.” (Scannella and Horner) Although this consensus continues to be debated to this day, this is not an unusual conclusion - dinosaurs do get reclassified from time to time.

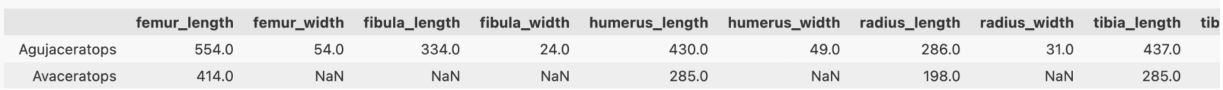
In this project, data mining techniques are used to explore the similarities in *Triceratops*, *Torosaurus*, and others genera to identify questionable classifications. We specifically focus on the *Ceraopsidae* family and hope to expand the analysis to others in the future.

Methods

The data for this project came from the Paleobiology Database (2022). Before retrieving tables, a filter was applied to select only records for the two main groups of dinosaurs, *Saurischia* (lizard-hipped) and *Ornithischia* (bird-hipped). Columns from three tables were selected for analysis:

- The **specimens** table returned 2688 records. Identification numbers, accepted names, and approximate age of the fossil specimens were selected for analysis.
- The **measurements** table returned 3152 records. Specimen identification number, measurement type (length, width, circumference, etc.), and measurement columns were used. There are more records in this table than the “specimens” table because some specimens had more than one measurement taken (for example, length and width of a femoral bone fossil).
- The **paleodb** table was used to collect taxa for all records and more detailed information for the *Ceratopsidae* fossils that were ultimately analyzed. This included geospatial data about where the fossil was discovered, such as latitude, longitude, and stratigraphy.

Once the data was collected, cleaned, and imputed (if possible), it needed to be structured in a way that allowed for comparison between different genera. Each one was represented as a vector of measurements for various bone fossils. Due to the small sample size, only one or two bone measurements existed for each genera. If more than one measurement was present, the mean value of all measurements were used. An example is shown in Figure 1.



	femur_length	femur_width	fibula_length	fibula_width	humerus_length	humerus_width	radius_length	radius_width	tibia_length	tib
Agujaceratops	554.0	54.0	334.0	24.0	430.0	49.0	286.0	31.0	437.0	
Avaceratops	414.0	NaN	NaN	NaN	285.0	NaN	198.0	NaN	285.0	

Figure 1: Snippets of the 1×21 vector representations of the genera *Agujaceratops* and *Avaceratops*

Then, comparisons could be made between pairs of vectors. To do this, some of the data had to be dropped if either genera was missing a measurement; comparisons were only made between bone measurements common to both. The number of common records was also recorded, along with cosine similarity and Euclidean distance for each pair.

$$cos(X, Y) = \frac{\sum_{i}^{p} x_i \cdot y_i}{\sqrt{\sum_{i}^{p} x_i^2} \cdot \sqrt{\sum_{i}^{n} y_i^2}}$$

Cosine Similarity

$$d(X, Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

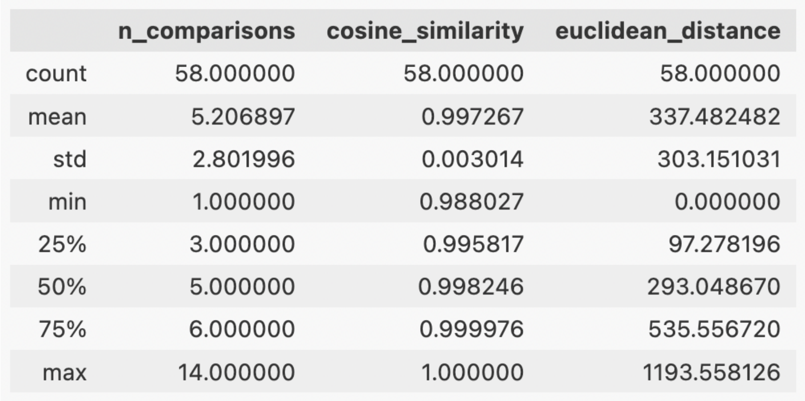
Euclidean Distance

Once these results were analyzed, the fossil discovery locations for the most similar pair was plotted on a map using Geopandas[1] according to Stewart (2018) to see where the most similar pair of *Ceratopsidae* genera were found.

[1] https://geopandas.org/en/stable/

Experimental Results

In total, 47 pairs of dinosaurs were compared with each other using cosine similarity and Euclidean distance and the results were analyzed. The summary of the results is shown in Figure 2.

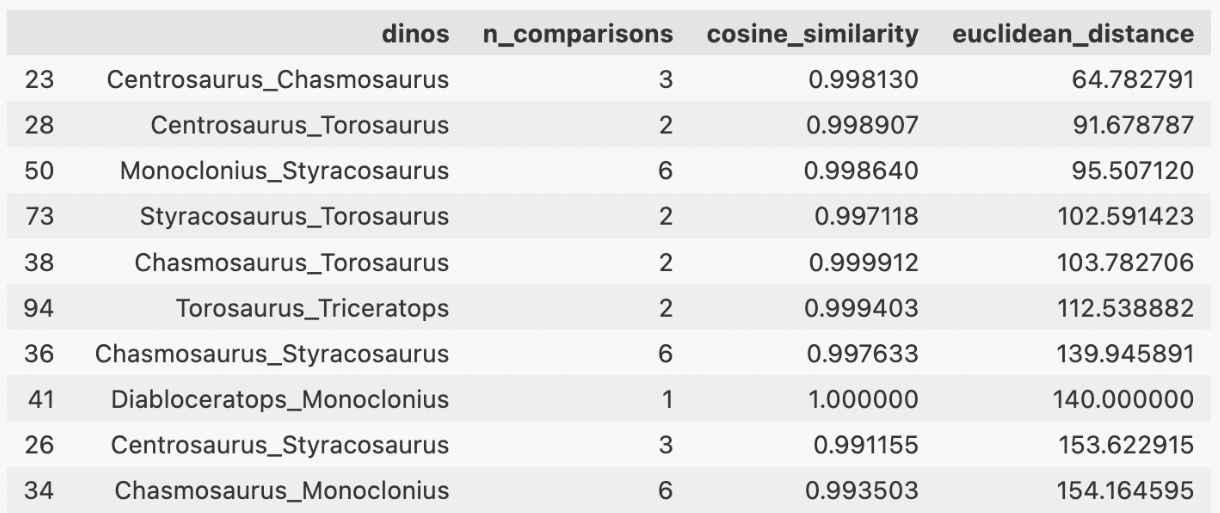


	n_comparisons	cosine_similarity	euclidean_distance
count	58.000000	58.000000	58.000000
mean	5.206897	0.997267	337.482482
std	2.801996	0.003014	303.151031
min	1.000000	0.988027	0.000000
25%	3.000000	0.995817	97.278196
50%	5.000000	0.998246	293.048670
75%	6.000000	0.999976	535.556720
max	14.000000	1.000000	1193.558126

Figure 2: Summary statistics for comparisons

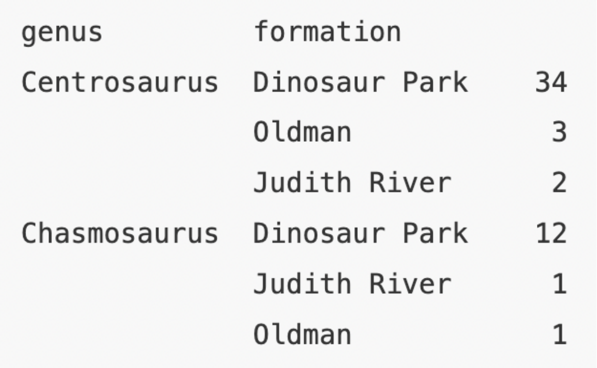
On average, 5 measurement comparisons were made between pairs of dinosaurs. Cosine similarity was not informative in this analysis, probably due to the small sample size. The results were then sorted in ascending order according to Euclidean distance (with the self comparisons dropped). Figure 3 shows a snapshot of the sorted list. Interestingly, the case of *Torosaurus* and *Triceratops* ranks highly (though only two measurements were compared). The closest pair according to Euclidean distance, the *Centrosaurus* and *Chasmosaurus*, were examined further.

It is important to note that just because two dinosaurs may look similar when comparing their bone structure, it doesn't mean much if they are deemed to have lived in different periods of time. Therefore, the next step was to compare the approximate age and location of the fossils. It turns out that both dinosaurs lived approximately 83.5-70.6 million years ago and were found in the same rock formations. Figure 4 shows the distribution of fossils in each formation.



	dinos	n_comparisons	cosine_similarity	euclidean_distance
23	Centrosaurus_Chasmosaurus	3	0.998130	64.782791
28	Centrosaurus_Torosaurus	2	0.998907	91.678787
50	Monoclonius_Styracosaurus	6	0.998640	95.507120
73	Styracosaurus_Torosaurus	2	0.997118	102.591423
38	Chasmosaurus_Torosaurus	2	0.999912	103.782706
94	Torosaurus_Triceratops	2	0.999403	112.538882
36	Chasmosaurus_Styracosaurus	6	0.997633	139.945891
41	Diabloceratops_Monoclonius	1	1.000000	140.000000
26	Centrosaurus_Styracosaurus	3	0.991155	153.622915
34	Chasmosaurus_Monoclonius	6	0.993503	154.164595

Figure 3: The top 10 “closest” pairs according to Euclidean distance.



```
genus          formation
Centrosaurus   Dinosaur Park    34
               Oldman            3
               Judith River      2
Chasmosaurus   Dinosaur Park    12
               Judith River      1
               Oldman            1
```

Figure 4: Fossil distribution in rock formations

Finally, Geopandas was used to create a plot of the latitude and longitude location of discovery of 40 *Centrosaurus* and 16 *Chasmosaurus* bone fossils. The results show significant overlap in the areas where they were found (see Figure 5).

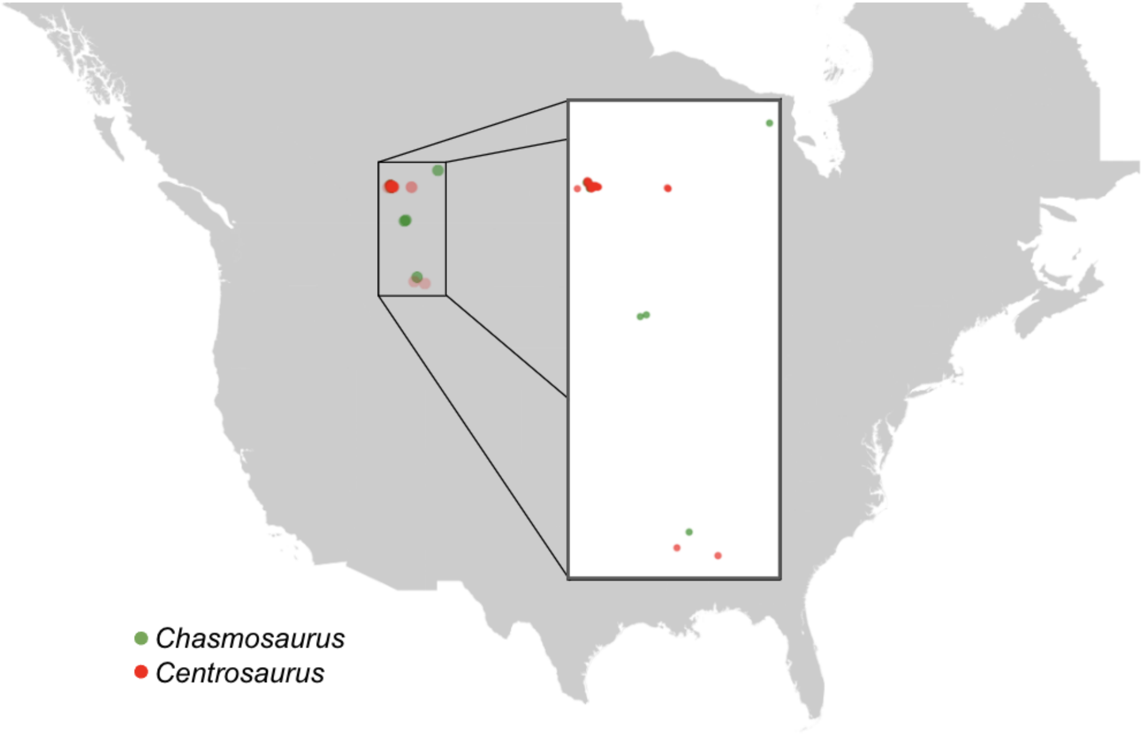


Figure 5: Map of discovery locations of *Chasmosaurus* and *Centrosaurus* bone fossils

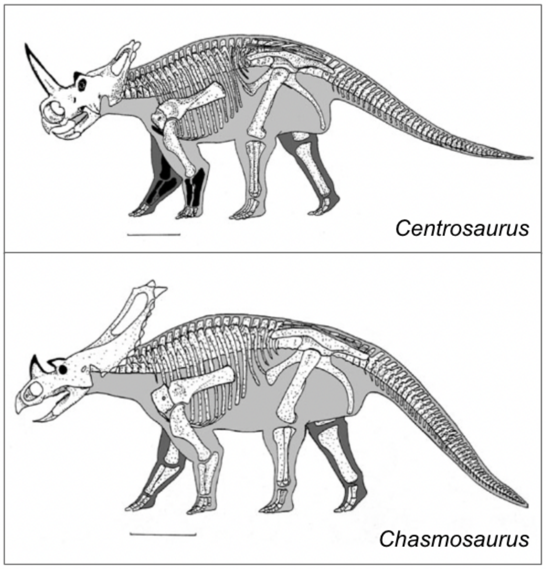


Figure 6: Artist’s renderings of *Centrosaurus* and *Chasmosaurus*

Conclusions

The two key findings in this work are:

1. Pairwise Euclidean distances between dinosaur genera can be used to call certain classifications into question.
2. Spatio-temporal features of the specimen records in question support the Euclidean distance metrics.

In other words, fossil pairs with lower (closer) Euclidean distances tend to be found in the same geographic regions and strata. This suggests that Euclidean distance might be a valid metric for finding pairs of dinosaur genera in the fossil record who might actually be the same.

Dinosaur fossil data is relatively rare, so while data mining techniques might not stand on their own as a classification tool, they may support paleobiologists who are faced with classifying newly discovered fossils or identifying problematic records.

Future Work

The sparse nature of bone fossil data limited this analysis. Traditional machine learning approaches, such as decision tree classifiers and k-Means clustering, struggled to find and make use of meaningful patterns. Also, a lot of data was lost because missing values were difficult or impossible to impute in a reliable way. Future iterations of this work should try to use a more robust dataset so that the Euclidean distance metric can be compared to others commonly used in classification.

The Paleobiology Database contains a wealth of information about body and trace fossils not just from dinosaurs, but from throughout the *Animalia*, *Plantae*, and *Fungi* kingdoms. The methods for calculating similarities and differences in this subset of data should be applied to broader categories to test for scale and validity. In addition, given the spatio-temporal nature of the fossil record, a time series analysis could be considered.

References

Hone, D. W. E., Farke, A. A., & Wedel, M. J. (2016). Ontogeny and the fossil record: what, if anything, is an adult dinosaur? *Biology Letters*, *12*(2). https://royalsocietypublishing.org/doi/10.1098/rsbl.2015.0947

The Paleobiology Database. (2022, November 28). *PBDB Data Service*. The Paleobiology Database. https://paleobiodb.org/classic/displayDownloadGenerator

Scannella, J. B., & Horner, J. R. (2010). Torosaurus Marsh, 1891, is Triceratops Marsh, 1889 (Ceratopsidae: Chasmosaurinae): synonymy through ontogeny. *Journal of Vertebrate Paleontology*, *30*(4), 1157-1168. https://www.tandfonline.com/doi/abs/10.1080/02724634.2010.483632

Stewart, R. (2018, October 31). *GeoPandas 101: Plot any data with a latitude and longitude on a map*. Towards Data Science. Retrieved December 10, 2022, from https://towardsdatascience.com/geopandas-101-plot-any-data-with-a-latitude-and-longitude-on-a-map-98e01944b972