# Creating And Embedding Spatio-Temporal Knowledge Graphs

Martin Böckling[1]

**Abstract**

An important paradigm of spatial analysis is the first law of geography, formulated by Tobler, which states that every geographic object is related to each other, but near geographic objects are more related than distant objects. Knowledge Graphs (KGs) in the spatial domain have received an increase in interest in the past years due to the possibility to realize the first law of geography. Within the paper, we outline the gaps present in the current research and sketch an approach on how to use KGs in the spatial domain by its involvement. In our outlined approach, we focus on the topological model Dimensionally Extended 9-Intersection Model (DE-9IM), which builds the base for the creation of our Spatio-Temporal Knowledge Graphs (STKGs). For the early-stage Ph.D., we aim to address different types of data preparation to investigate the influence of the different data preparations. Furthermore, the idea and possibilities to explore a specialized embedding methodology for the spatio-temporal domain in more depth are outlined. In a preliminary experiment for predicting wildfires, we demonstrate the impact of the formulated research approach and provide an outlook to further research possibilities within the area of STKGs.

**Keywords**

Spatio-Temporal Knowledge Graph, Knowledge Graph embeddings, Wildfire prediction

## 1. Introduction

KGs are a widespread data representation to not only model knowledge representable, but also allow modeling complex systems together with the description of the relationship (predicates). A KG in general consists of a triple (subject, predicate, object) [1]. Within the area of spatial analysis, complex systems play a vital role in modelling reality into data representations. Tobler's first geographic law states, that "Everything is related to everything else, but near things are more related than distant things." [2]. This simplified rule finds representation in spatial interpolation or spatial auto correlation analysis and is a major aspect for modeling complex systems in the spatial domain and is still considered to be applicable [3].

KGs can therefore serve within the spatial analysis domain an important role, representing complex systems within the spatial domain. To make KGs applicable for machine learning tasks in the spatial domain, the respective KG needs to be present in a numeric representation. In the domain of spatial data, often the temporal component plays a key role when predicting spatial events like wildfires. Due to the dynamic changes present in the data, static embedding methodologies come with a shortcoming of taking the dynamic changes into account. Furthermore,

*22nd International Semantic Web Conference, November 06–10, 2023, Athens, Greece*
martin.boeckling.gast@uni-mannheim.de (M. Böckling)
https://github.com/MartinBoeckling (M. Böckling)
0000-0002-1143-4686 (M. Böckling)

in general, a methodology of a STKG is not yet present which takes the Tobler's first law into account. Especially, the inner relationships between geographic geometries and geographic grid cells are, to the best of our knowledge, not yet considered in KG schemas.

This early-stage PhD focuses on the exploration and application of STKGs in machine learning tasks. Together with an exploration of possible alternative embedding approaches and a framework of a STKG creation, the PhD analyzes its use to tackle the current gaps in the research domain. It is shown in preliminary results that an involvement of STKG embeddings improves the overall prediction task.

## 2. State of the Art

Spatial data provides oftentimes the challenge that diverse datasets need to be unified for prediction. To overcome the challenges faced with spatial data, different options are available for the unification of diverse datasets. One frequent used methodology is the use of a spatial grid, where each single spatial grid cell has an assigned unique Identifier (ID) [4]. The involvement of spatial grids associates the data specifically to the individual grid cell. Current approaches do not account for neighboring relations, which have a potential influence and are not captured in the dataset. As another option, research focuses not on individual grid cells but assigns features which are within a predefined distance to a point of interest or calculates the distance to a feature of interest [5]. This ignores features which are slightly above the predefined distance of a point of interest, or does not relate features which are spatially near to each other.

In the domain of spatial data, different data representations for graph-based datasets are possible. One possibility of such graph data representation are static node and edge pairs. An example of static graphs are road networks or electricity grids [6]. The possibility to connect spatial information via a graph allows modeling links between one single geometry or multiple geometries to represent complex systems [7].

Different research papers have looked at the possibility to model KGs based on spatial information. In comparison to spatial graphs, the edges within a KG depict the relationship [8].

For the area of KGs within the spatial domain, we focus on a selection of different KGs. One KG capturing OpenStreetMap (OSM) data is called WorldKG. The KG is structured by transposing the features from OSM by relating different categories. Within the KG, tags derived from OSM serve as parent nodes within the KG WorldKG. Individual features from OSM are exposed to individual nodes. Additionally, the geometries of each OSM entity are also exposed as nodes within WorldKG [9].

KnowWhereGraph builds up on a variety of different datasets covering hazard information, climate data, soil properties, crop and land-cover types, demographics and human health. For the integration of the different datasets, the S2 discrete hierarchical grid is used to unify the location data from different sources. Each grid cell within the discrete hierarchical grid serves as a unique ID identifying the region. Together with the grid cell ID different other regional information are mapped to the area like ZIP codes, Administrative Regions or Climate Division Boundary. [10]

Specifically for OSM data, osm2rdf provides a converter which transforms the geographical data into a KG. The proposed approach provides an efficient possibility to transform OSM data

together with a SPARQL endpoint to consume the most recent snapshot from OSM. Every week, the latest snapshot derived from OSM replaces the snapshot from the previous week. For each entity within the KG the OSM ID is used as a unique identifier. [11]

An approach using embeddings for STKGs has been introduced with the method called ST-NewDE, which is based on Dihedron Algebra to calculate the embeddings for the STKG. The ST-NewDE approach uses for the calculation of the embeddings quintuples to expand the general triple structure with a temporal and spatial entity. The considered spatial information and temporal information showed better performance than already existing methodologies [12]. From the spatial research the involvement of Spatio-Temporal Graph Neural Networks has been most prominently used on traffic forecasting datasets. Current approaches combine the temporal information of a dynamic graph together with the temporal dimensions. The different architectures currently learn the temporal and spatial dimension separately to then combine the representations together for the overall learning task [13].

The current approaches within the domain of spatial analysis are limited regarding an accurate representation of Tobler's first law of geography. Especially, the relationships between geometries and the inclusion of neighboring effects between geometries is limited and not explored in the above presented works. All presented works do not model neighborhood relations in the knowledge graph and therefore limit their semantics to the locality of geometries. Furthermore, the temporal dimension of KG is not considered in the presented works. In section 3 the problem statement that the PhD thesis tackles is outlined.

## 3. Problem statement and Contributions

For the current approaches outlined in section 2, the relationship between different geometries within a defined space are not depicted. Within the geographic domain, the influence of relationships between geometries has shown to have beneficial effects to model the general relationship between geometries [14]. A framework for the modeling of relationships within the spatial area is the DE-9IM methodology. It relies on a comparison between two geometries and the extraction of a relationship determination between the two geometries. The relationship of the DE-9IM is derived from a nine field intersection matrix, that intersects the interior, boundary and exterior of each geometry to the other geometry. These patterns capture the possible relationships between two spatial objects, such as points, lines, or polygons, based on their relative positions and spatial overlaps. It becomes possible to analyze and reason about the spatial relationships between objects, such as determining if a point is inside a polygon or if two polygons intersect each other [15]. The incorporation of relationships between single geometries, geometries, and grid cells and between grid cells helps to model a more accurate representation of reality. The problem of only looking at stationary effects limits the possibility to follow the geographic first law formulated by Tobler [2].

Furthermore, current KGs in the spatial area do not capture changes continuously over a time period. As within the spatial area, especially in urban areas, different changes can happen frequently. In many KGs, only a limited possibility to capture periodically changes is implemented within the current approaches. Furthermore, only KnowWhereGraph uses with S2 a discrete global hierarchical grid to mark the position of a spatial feature within the spatial

KG.

To answer the problem statements described in the above paragraphs, the PhD thesis is trying to answer the following research questions:

RQ1. Do different discrete hierarchical grids KGs influence embeddings besides a hexagonal discrete hierarchical grid?

RQ2. Can within a KG the embedding quality be improved by using spatial weights which assign higher weights for nearer geometries?

RQ3. Can the splitting of the temporal and spatial information help to capture dynamic changes?

## 4. Research Methodology and Approach

As a base for the PhD thesis, the creation of a general STKG approach is necessary to evaluate the effectiveness of upcoming proposed methodologies. Due to the open-source availability of geographic data, OSM plays an important role in the creation of an STKG. The OSM dataset that can be derived consists of geometric entities like Point geometries, LineString geometries or Polygon geometries. Furthermore, OSM contains relational information like the speed limit for a specific street. As a base for the STKG the h3 discrete global grid will be used to divide regions of interest into fixed spatial grid cells. By using the h3 grid, a general extension of the created KG is possible. To make the use of the STKG possible in Spatio-Temporal use cases, the framework can capture dynamic data from OSM. For the research and solving of the formulated problem in chapter 3 the geometries of a geographic object are related to a spatial grid by using the geometric topology with the DE-9IM method. Based on the determined relationship between the spatial object and the spatial grid, a KG is constructed which depicts the relationship between two geographic geometries. This allows the possibility for a precise modeling of the geographic situation.

In a second step, after the creation of a STKG the existing embedding technologies used for KGs are compared regarding its performance and applicability for machine learning tasks. If needed, a revision of already existing architectures is proposed by combining methodologies from the spatial and the semantic web technology domain.

A possible direction for implementing spatial capabilities into existing embedding methodologies are distance-based measurements to weight for each node within a KG the respective nodes. The weighting of more distant nodes within the network follows Tobler's first law of geography, which would give more weight to nodes nearer to the current node. The overall approach is implemented in the spatial domain by interpolation techniques like Inverse Distance Weighting (IDW) and Kriging. As the proposed architecture of the STKG captures the dynamic changes over time, a transformation of already existing static embedding methodologies will be researched. Furthermore, it is tested if the splitting of temporal embedding creation and structural embedding creation helps to capture the dynamic elements in the STKG.

As displayed in figure 1, the plan for the early-stage PhD is to tackle the mentioned aspects over the research period and evaluate them properly. In the first year of the PhD, the general framework on the modeling of a STKG is planned to put into state and evaluated on different

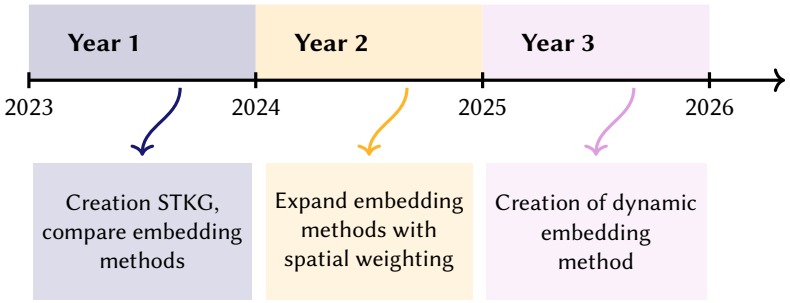


**Figure 1:** Proposed timeline of PhD research

datasets with existing embedding methodologies and data preparation techniques. Furthermore, the different embedding methodologies present in the research area are compared to the selected datasets regarding performance. For the second year of the PhD, the goal is to look at existing embedding methodologies and expand if possible the methods to allow a distance-based weighting for the embedding generation. In the third year of the PhD, the goal is to focus on the possibility to capture dynamic changes without need of embedding alignments. Currently, embedding methodologies for KGs rely on static KGs, which requires the need to align the generated embeddings or retrain the embeddings. As our research is still at the beginning, the proposed timeline is still subject to change.

## 5. Evaluation Plan

The evaluation for the embedding methodologies is assessing, whether the constructed embedding methodology outperforms the current state-of-the-art embedding approaches. It is based on prediction tasks related to Spatio-Temporal prediction tasks. We differentiate in the overall evaluation between three configurations, with the base case, the hybrid case and the network case. The Base Case dataset table-based prediction task with defined variables from the Data Preparation. The Hybrid Case dataset combines table-based structure together with created embeddings of the modeled OSM STKG. The Network Case dataset unifies all variables into the STKG and create embeddings. The embeddings are joined with the labels/ predictor variables together.

To join the created embeddings together with the labels and data from the training datasets, the extracted grid IDs from the h3 grid is derived. For the different datasets and learning tasks, the according measurements for the overall quality are compared to each other. This includes for classification tasks the F1-score and for regression tasks the Root Mean Squared Error (RMSE) as main evaluation metrics. To make an assessment if a certain methodology performed better in a specific dataset constellation, the significance intervals between the different datasets is determined. If the significance intervals do not overlap, the result is considered to be significant compared to the other result.

For the PhD research, the plan is to evaluate the outlined approach on a selection of different datasets. The involved datasets capture in all scenarios geographic information with separate influence variables, ranging from classification to regression tasks. In section 6 the initial results

derived from an investigated use case are outlined based on the described approach in section 4.

## 6. Preliminary results

The results and conclusion showcased in this section are early-stage results of the PhD research and are limited to a selection of OSM data for the territory of California. As the overall field of STKG and its embeddings is a new research field, the focus on the first research result was on the validation of the overall approach for the STKG creation.

For the setup of the first prediction, a STKG for a wildfire prediction for the territory of California is constructed. The prediction covers a period from 2010 to 2021, in which the data is present on a monthly basis. For the data constellation of the wildfire prediction four datasets are used: Wildfire data [16], weather data [17], landscape data [18], OSM data [19]. Besides the wildfire data and the OSM data, the weather data and landscape data are used for the wildfire prediction. As mentioned in the introduction of this section, a selection of available OSM tags are derived from the OSM API.[1]

For the base case dataset, the four different datasets described in the above paragraph are unified into a tabular data structure by using a spatial grid. The spatial grid consists of hexagonal grid cells and covers in total the territory of California. Each hexagonal grid cell has a size of 20 $km^2$. After joining the different datasets to the respective grid cell, the dataset can be used to predict wildfires on influence factors derived from the datasets used. Each row within the dataset is a unique combination between a specific month and a spatial grid cell. For the overall assessment on the performance, each dataset constellation uses the period of 2020 to 2021 as a test dataset and the period of 2010 to 2019 as a train dataset.

For the creation of the STKG the creation is divided into two logical parts. The first part is the conversion of tabular data into a KG. As a starting point of the conversion, the spatial grid cell builds the base to orient the data within the KG. The column which stores the spatial grid cell is used together with the month as a long orientation column to pivot the columns stored in the base dataset. Therefore, for each row within the tabular dataset, each column is converted to a new row. The column names are used as predicates within the KG and the values stored in the column are used as objects within the KG. The ID of the spatial grid cell is the subject in the KG.

For the second part of the KG creation, the geographic geometries derived from OSM are related to all individual grid cell using the DE-9IM method. Each grid cell furthermore is related to each other using the DE-9IM method. The described KG creation follows the outlined approach in section 4 and relates the OSM geometries to the created Spatial Grid. This allows to not only relate all OSM data to one grid cell, but also relate the neighboring information to the grid cell currently creating the embedding for. As an embedding methodology, the standard RDF2Vec approach is used. To prevent that between the different periods a different vector space is constructed, a partial embedding alignment is performed, which is based on the vector alignment of HistWord [20]. It is important to state that we assume that more specific embedding methodologies would improve the overall prediction performance.

[1]Coding of experiment can be found here: github.com/MartinBoeckling/\WildfirePredictionSTKG.

For the wildfire prediction, the Extreme Gradient Boosting (XGBoost) algorithm is used [21]. For the selection of the hyperparameter values, a Bayesian optimization is used to select the optimal parameter values for XGBoost. The selected hyperparameters are derived from the recommendations of tunable hyperparameters. In the following, the preliminary results are compared using the calculated significance intervals for the derived F1 score. The Base Case dataset has achieved an F1-score of 0.3478±0.0010 with only using the table-based data structure. The Hybrid Case dataset has achieved an F1 score of **0.3803**±0.0011 by combining the table-based data with the vectors from OSM. The network case dataset with the unification of all variables into a STKG achieved a F1 score of 0.0107±0.0002, which is significantly lower than the Hybrid Case dataset. In section 7, the different results are put into context and briefly analyzed.

## 7. Lessons learned

Within this paper, the preliminary results of the starting point of the PhD research are displayed. For the Hybrid Case dataset, the wildfire prediction showed a significant better performance compared to the Base Case dataset. However, the Network Case dataset showed a significant lower performance compared to the Hybrid Case or Base Case dataset. The result for the Network Case dataset could be improved by using a SMOTE oversampling instead of the Random Oversampling. Based on the initial results derived from the wildfire prediction, it has been shown that the inclusion of KG embeddings helped to improve the detection of wildfires. Nevertheless, the dataset constellation in which every dataset is used for the STKG showed a significant lower performance than the other dataset constellations. Whether the embedding methodology or the KG construction are the root cause of the significant low performance needs to be investigated.

The initial results of our research show promise to further investigate the topic of STKG and the creation of embedding structures. Based on the geographic first law by Tobler, we believe that the outlined approach has the potential to model more accurately complex systems within the Spatio-Temporal domain and pose the possibility to significantly improve the performance of different prediction tasks in the spatial domain.

## References


[1] L. Wang, T. Ai, The comparison of drainage network extraction between square and hexagonal grid-based dem, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-4 (2018) 687–692.

[2] W. R. Tobler, A computer movie simulating urban growth in the detroit region, Economic Geography 46 (1970) 234–240.

[3] N. Manning, Y. Li, J. Liu, Broader applicability of the metacoupling framework than tobler's first law of geography for global sustainability: A systematic review, Geography and Sustainability 4 (2023) 6–18. doi:10.1016/j.geosus.2022.11.003.

[4] P. Rigaux, M. Scholl, A. Voisard, Spatial Databases: With application to GIS, 2 ed., Morgan Kaufmann Publishers, San Francisco, 2002.

[5] S. J. Kim, et al., Multi-temporal analysis of forest fire probability using socio-economic and environmental variables, Remote Sensing 11 (2019).
[6] M. Barthélemy, Spatial networks, Physics Reports 499 (2011) 1–101.
[7] R. G. Morris, M. Barthelemy, Transport on coupled spatial networks, Physical Review Letters 109 (2012) 9–13.
[8] J. Wang, X. Wang, C. Ma, L. Kou, A survey on the development status and application prospects of knowledge graph in smart grids, IET Generation, Transmission & Distribution 15 (2021) 383–407.
[9] A. Dsouza, N. Tempelmeier, R. Yu, S. Gottschalk, E. Demidova, Worldkg: A world-scale geographic knowledge graph, in: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, ACM, 2021.
[10] K. Janowicz, et al., Know, know where, knowwheregraph: A densely connected, cross–domain knowledge graph and geo–enrichment service stack for applications in environmental intelligence, AI Magazine 43 (2022) 30–39.
[11] H. Bast, P. Brosi, J. Kalmbach, A. Lehmann, An efficient rdf converter and sparql endpoint for the complete openstreetmap data, in: X. Meng (Ed.), 29th ACM SIGSPATIAL, The Association for Computing Machinery, Inc, New York, NY, 2021, pp. 536–539.
[12] M. Nayyeri, et al., Dihedron algebraic embeddings for spatio-temporal knowledge graph completion, in: P. Groth, et al. (Eds.), The Semantic Web, volume 13261 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022, pp. 253–269.
[13] K.-H. N. Bui, J. Cho, H. Yi, Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues, Applied Intelligence 52 (2022) 2763–2774.
[14] J. Chen, C. Li, Z. Li, C. Gold, A voronoi-based 9-intersection model for spatial relations, International Journal of Geographical Information Science 15 (2001) 201–220.
[15] E. Clementini, P. Di Felice, P. van Oosterom, A small set of formal topological relationships suitable for end-user interaction, in: D. Abel, B. C. Ooi (Eds.), Advances in Spatial Databases, volume 692 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg and Imprint: Springer, Singapore, 1993, pp. 277–295.
[16] J. Ruiz, J. Lázaro, I. Cano, P. Leal, Burned area mapping in the north american boreal forest using terra-modis ltdr (2001–2011): A comparison with the mcd45a1, mcd64a1 and ba geoland-2 products, Remote Sensing 6 (2014) 815–840.
[17] M. J. Menne, I. Durre, R. S. Vose, B. E. Gleason, T. G. Houston, An overview of the global historical climatology network-daily database, Journal of Atmospheric and Oceanic Technology 29 (2012) 897–910.
[18] L. Yang, et al., A new generation of the united states national land cover database: Requirements, research priorities, design, and implementation strategies, ISPRS 146 (2018) 108–123.
[19] OpenStreetMap contributors, Plante dump, 2017. URL: https://www.openstreetmap.org.
[20] W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2016, pp. 1489–1501.
[21] T. Chen, C. Guestrin, Xgboost, in: B. Krishnapuram (Ed.), KDD'16, Association for Computing Machinery, New York, New York, 2016, pp. 785–794.