# Extraction of terms highly associated with named rivers

**ICEI 2018: 10th International Conference on Ecological Informatics**

**Session S1.6: Semantics for Biodiversity and Ecosystem Research**

Jena, Germany, 24-29 September 2018

Universidad de Granada


Juan Rojas-García
juanrojas@ugr.es

Riza Batista-Navarro
riza.batista@manchester.ac.uk

Pamela Faber
pfaber@ugr.es


Definition

River: large stream of water flowing in a bed or channel and emptying into the ocean, a sea, a lake, or another stream.

Terms

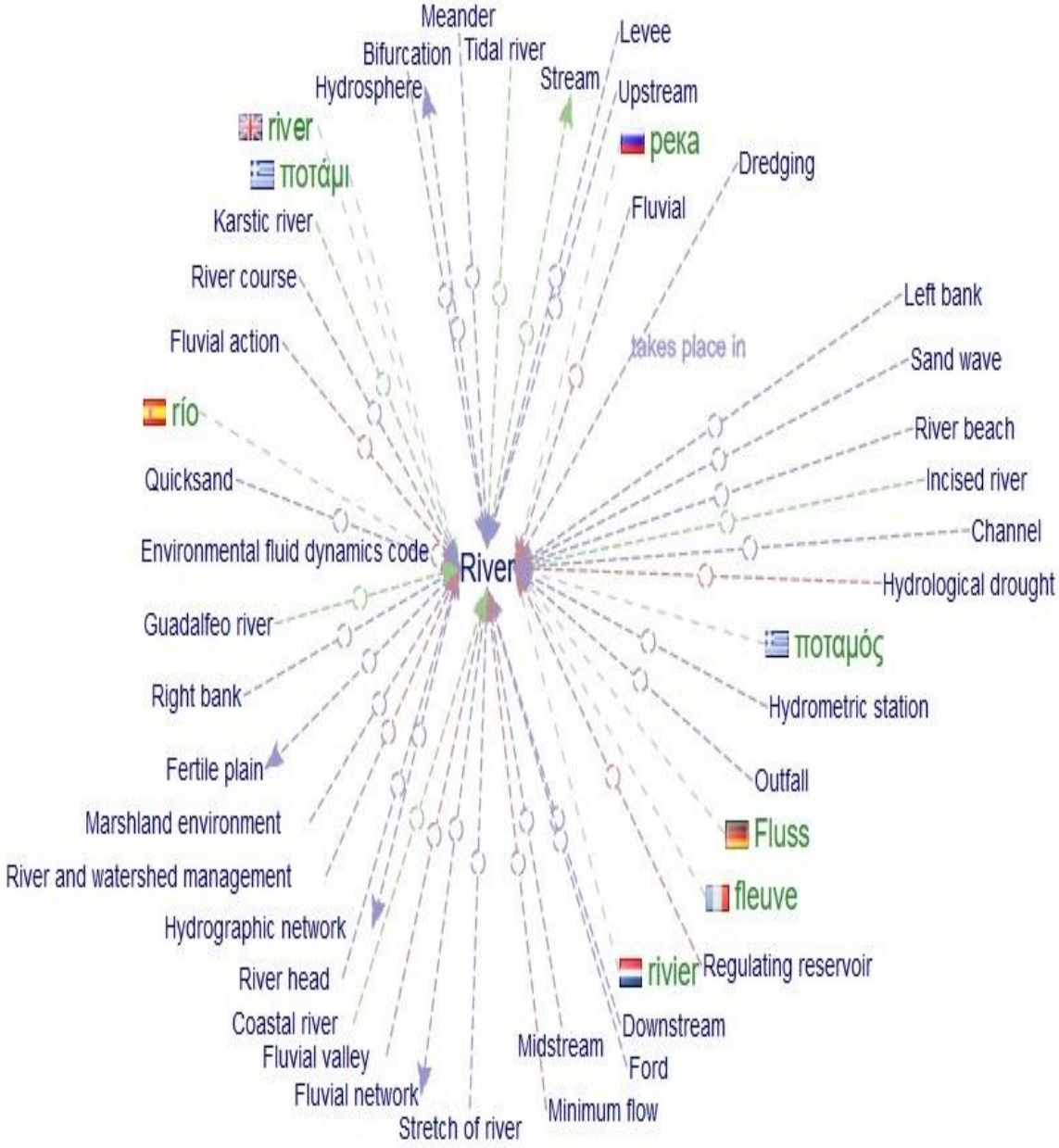
river
río
Fluss
река
fleuve
rivier
ποταμός
ποτάμι

Resources

Conceptual categories

Phraseology

History | Search results | A-Z | Path | Search concordances



# 1.- Introduction

**Motivation**:

- EcoLexicon permits the contextualization of data so that they are more relevant to specific:
  - subdomains of knowledge,
  - communicative situations, and
  - **geographic areas**

[*León Araúz et al. 2013*]

# 1.- Introduction

**Motivation**:

- However, to facilitate the **geographic contextualization** of concepts such as those belonging to the semantic category of **LANDFORM**, it is necessary to know:
  - **what terms are related to** each type of landform
  - and **how the terms are related to each other**

# 1.- Introduction

**Motivation**:

- In order to extract the terms related to **named rivers** (e.g. *Mississippi River*, *Nile River*, etc) from a specialized corpus of research papers, we rely on semi-automatic methods based on **Distributional Semantic Models** (DSMs).

# 1.- Introduction

**Motivation**:

- In this task, we face two issues:
  1. Corpus-based lexical studies on specialized domains and for specific purposes normally rely on **small, specialized corpora**, which, in the case of written ones, range from 250.000 to around 6 million tokens [*Flowerdew, 2004: 19*; *O'Keeffe et al., 2007: 4*].
  2. The performance of DSMs for the extraction of knowledge has been extensively evaluated in very **large, general corpora**, but **not** in **small, specialized corpora** [*Bullinaria & Levy, 2007, 2012*; *Baroni et al., 2014*; *Kiela & Clark, 2014*; *Lapesa et al., 2014*].

# 1.- Introduction

**Objective:**

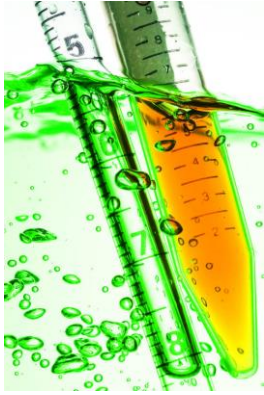
- The aim of this paper was to look for **parameter combinations of DSMs**, suitable for the extraction of three specific **semantic relations** held by **named rivers**, namely, *takes_place_in*, *causes*, and *located_at*.
- For that purpose, an experiment was carried out, in which different DSMs were built on a **small specialized corpus**, and then **evaluated on gold standard data** manually extracted from the same corpus.

# 1.- Introduction

## Distributional Semantic Model (DSM)

- A **DSM** can be a very useful tool for terminology, as it can help identify **semantic relations** between terms based on corpus data [*Bernier-Colborne & Drouin, 2016*]
- A DSM produces **vector representations of words**, based on the contexts in which they appear in a corpus, the underlying hypothesis being that words that appear in similar contexts have similar meaning [*Harris, 1954; Firth, 1957*]

# 2.- Materials



- **EcoLexicon English Subcorpus** on the domain of Coastal Engineering [http://manila.ugr.es/visual/index_en.html]
  - ➢ around 7 million word tokens
  - ➢ Now publicly available in Open corpora (Sketch Engine)
  - ➢ https://the.sketchengine.co.uk/open/

LexiCon research group

**Open corpora on Sketch Engine: EcoLexicon**

November 30, 2017 MichaelHB corpora, tools corpora, tools

If you work with or read about corpora, you are probably familiar with Sketch Engine. If you aren't familiar with it, it is described on its own website as "the ultimate corpus tool", and that's maybe not an exaggeration. You can do a ton of cool stuff with it. Sketch Engine also provides access to hundreds of ready-to-use corpora in close to a hundred languages.

However, it requires a subscription (although a 30-day trial is available for starters). This puts some people off from it; after all, there are a lot of free resources out there.

What some people may not realize, though, is that there are some "open corpora" on Sketch Engine that can be explored (with all of Sketch Engine's features) without registration.

Sketch Engine anonymous

Sketch Engine

Home
Register
Log in
Lost password

Support

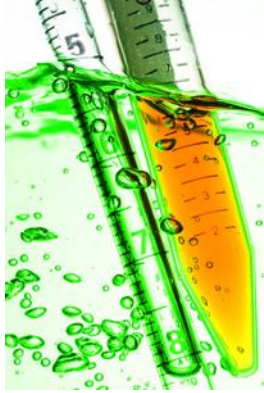
User guide
Contact support

**Open corpora**

| Language | Corpus name | Words |
|---|---|---|
| Chinese Simplified | Guangwai - Lancaster Chinese Learner Corpus | 1,289,065 |
| English | ACL Anthology Reference Corpus (ARC) | 62,196,334 |
| English | British Academic Spoken English Corpus (BASE) | 1,186,290 |
| English | British Academic Written English Corpus (BAWE) | 6,968,089 |
| English | Brown | 1,007,299 |
| English | EcoLexicon English (Environment) | 23,169,446 |
| N'Ko | Corpus Nko | 3,803,556 |

Get more corpora by registering for an account. See overview of available corpora.

# 2.- Materials



## Corpus Design:

• This subcorpus on Coastal Engineering is integral part of the **English EcoLexicon corpus**, which currently contains over 59 million words in English and is focused on the **environmental domain**

• It was **manually compiled** for the development of **EcoLexicon** [http://ecolexicon.ugr.es], an **electronic, multilingual, terminological knowledge base** on environmental sciences

• To maximize representativeness, the corpus was designed based on criteria proposed by *Sinclair (1991, 2005)*, *Meyer (2004)*, and *Biber (2008)*: **balance**, **diversity of sources**, **availability of texts in electronic form**, **period**, **size**, **use of complete texts**, and **variety of writers**

# 2.- Materials

- Use of different **Distributional Semantic Models** to represent the named rivers as vectors of co-occurrence frequencies.
- Use of **three gold standard datasets** for the semantic relations
  - *takes_place_in*,
  - *causes*, and
  - *located_at*.

# 2.- Materials

- Examples of the **gold standard dataset** for the semantic relation *takes_place_in*:
  - Consolidation of the land *takes_place_in* Mississippi river mouth
  - Runoff *takes_place_in* Mississippi river basin
  - Sea level rise *takes_place_in* Mississippi river delta

- **Process** *takes_place_in* **named_river**
- Example from the corpus:
  - ✓ *... **Consolidation of the land** is occurring, as noted before, at the **mouth of the Mississippi River**, where the …*

# 2.- Materials

- Examples of the gold standard dataset for the semantic relation *causes*:
  - Mississippi river *causes* Saint Bernard river delta
  - Mississippi river *causes* soft mud
  - Mississippi river *causes* sediment transport

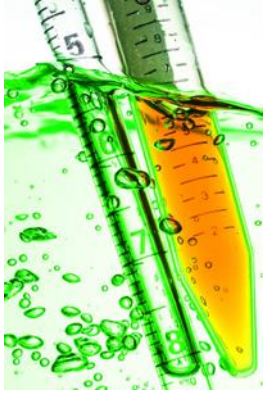
- **Named_river** *causes* **process / entity**
- Example from the corpus:
  - ✓ *... The Chandeleurs Islands are remnants of the **Saint Bernard River delta**, formed by the **Mississippi River**.*

# 2.- Materials

- Examples of the gold standard dataset for the semantic relation *located_at*:
  - Jetty *located_at* Mississippi river mouth
  - Soft mud *located_at* Mississippi river mouth
  - Barrier island *located_at* Mississippi river mouth

- **Entity** *located_at* **named_river**
- Example from the corpus:
  - ✓ *... **barrier islands**, <u>located</u> near the **mouth of the Mississippi River**.*
  - ✓ *... at the **mouth of the Mississippi River**, where the **soft muds** deposited by the Mississippi River are consolidating.*

# 2.- Materials



- ***R programming language*** for statistical analysis [*R Core Team, 2017*]

- R package ***wordVector*** [*Schmidt & Li, 2016*] for the prediction-based models *word2vec*

- R package ***quanteda*** [*Benoit, 2018*] for the count-based models

- **GeoNames** database to automatically match the designations of named rivers in the corpus [http://www.geonames.org/search.html? ]

# 3.- Method

## Pre-processing:

1. Normalization of the corpus
   - Cleaning up
   - Tokenized, POS-tagged and Lemmatized
   - **MWTs from EcoLexicon were matched in the corpus and joined with underscores**
3. Punctuation marks, number and symbols were removed
4. Character strings with **less than 3 characters** were removed
5. Function words (closed words such as prepositions, determiners etc.) were removed

# 3.- Method

## Parameters evaluated:

## Count-based models

- **Size of the context window: 1-10 words**
- **Weighting scheme:**
  - **log-likelihood** (frecuently used in Computational Linguistics)
  - **Positive Pointwise Mutual Information (PPMI)**
  - **t-score**
  - **z-score** **(**frecuently used in Computational Lexicography**)**

# 3.- Method

## Parameters evaluated:

## Prediction-based models (*word2vec*)

- **Architecture: CBOW or skip-gram**
- **Negative samples: 5, 10 or none (hierarchical softmax)**
- **Subsampling threshold: low ($10^{-5}$), high ($10^{-3}$), none**
- **Size of context window: 1-10 words**
- **Dimensionality of word embeddings: 100 or 300**

# 3.- Method

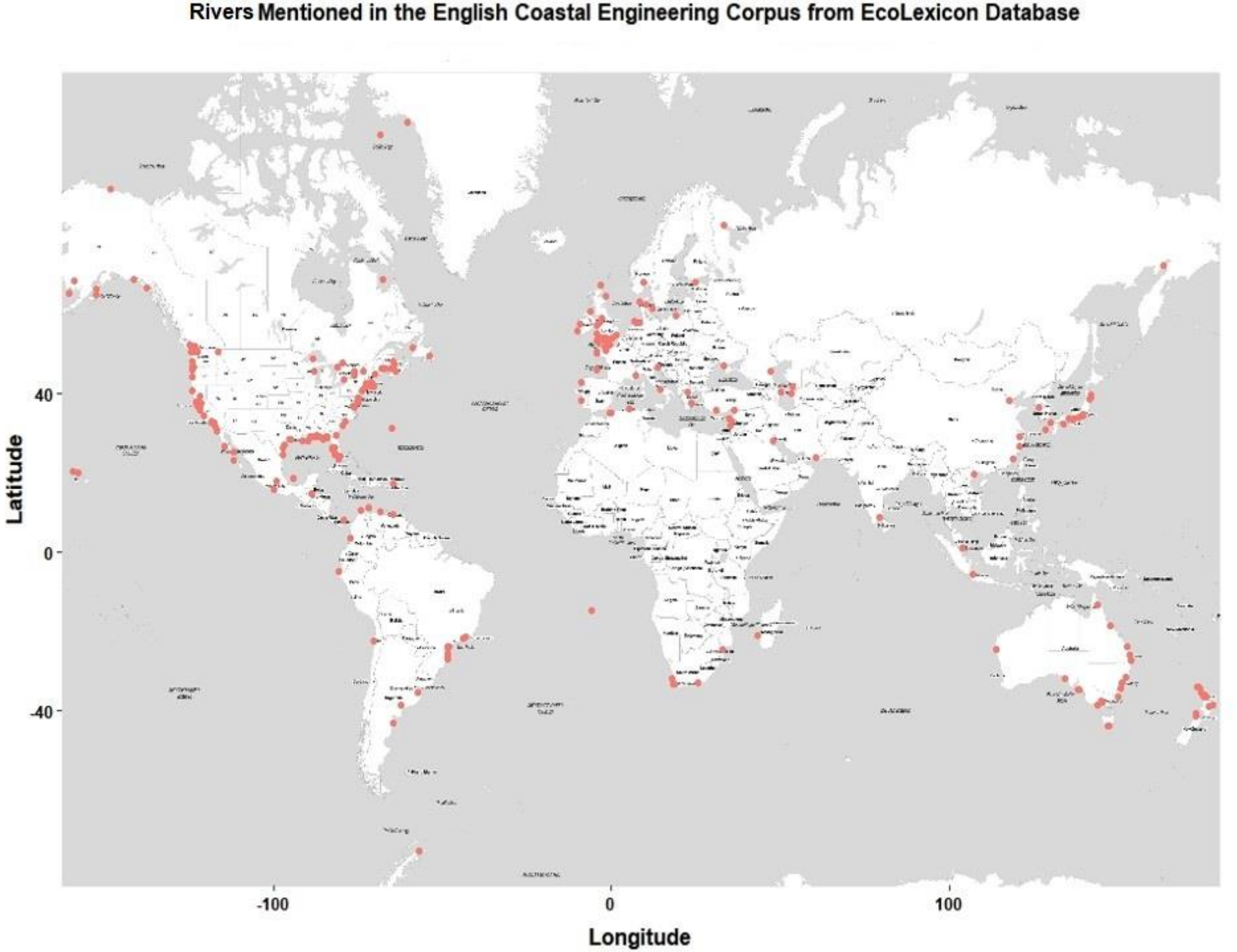
## Evaluation:

- The measure used to evaluate the models is **Mean Average Precision** (MAP).
  - **MAP** tells how accurate the sorted list of neighbours we get for a given query is, based on the rank of its related terms according to the gold standards.
- The **cosine similarity** is used to measure the similarity between the term vectors.

# 3.- Method

## Matching of named rivers in the corpus:

- The designations of the named rivers mentioned in the whole corpus were automatically matched by means of the named rivers stored in **GeoNames** database dump.
- **320** named rivers were recognized in the corpus.


Rivers Mentioned in the English Coastal Engineering Corpus from EcoLexicon Database
Latitude
40
0
-40
-100
0
100
Longitude

# 4.- Results

## Comparing count-based and prediction-based models

- We compare:
  - Bag of words (**BOW**): count-based model
  - *word2vec* (**W2V**): prediction-based model

by observing the MAP of each model on each dataset.

- The **maximum MAP** (with average and standard deviation in brackets) is shown.
- The **BOW** model achieves a higher MAP than **W2V** on the three semantic relations if its parameters are tuned correctly.

| Dataset | BOW | W2V |
|---|---|---|
| takes_place_in | **0.544** (0.347 ± 0.118) | 0.346 (0.298 ± 0.042) |
| located_at | **0.418** (0.321 ± 0.056) | 0.221 (0.196 ± 0.013) |
| causes | **0.383** (0.247 ± 0.055) | 0.199 (0.153 ± 0.019) |

# 4.- Results

## Comparing count-based and prediction-based models

- The **maximum MAP** of BOW model on the three datasets is achieved when:
  - The statistical association measure is **log-likelihood** for the 3 semantic relations
  - The window size for the relation ***takes_place_in*** is: **5** words
  - The window size for the relation ***causes*** is: **3** words
  - The window size for the relation ***located_at*** is: **2** words

| Dataset | BOW | W2V |
|---|---|---|
| takes_place_in | **0.544** (0.347 ± 0.118) | 0.346 (0.298 ± 0.042) |
| located_at | **0.418** (0.321 ± 0.056) | 0.221 (0.196 ± 0.013) |
| causes | **0.383** (0.247 ± 0.055) | 0.199 (0.153 ± 0.019) |

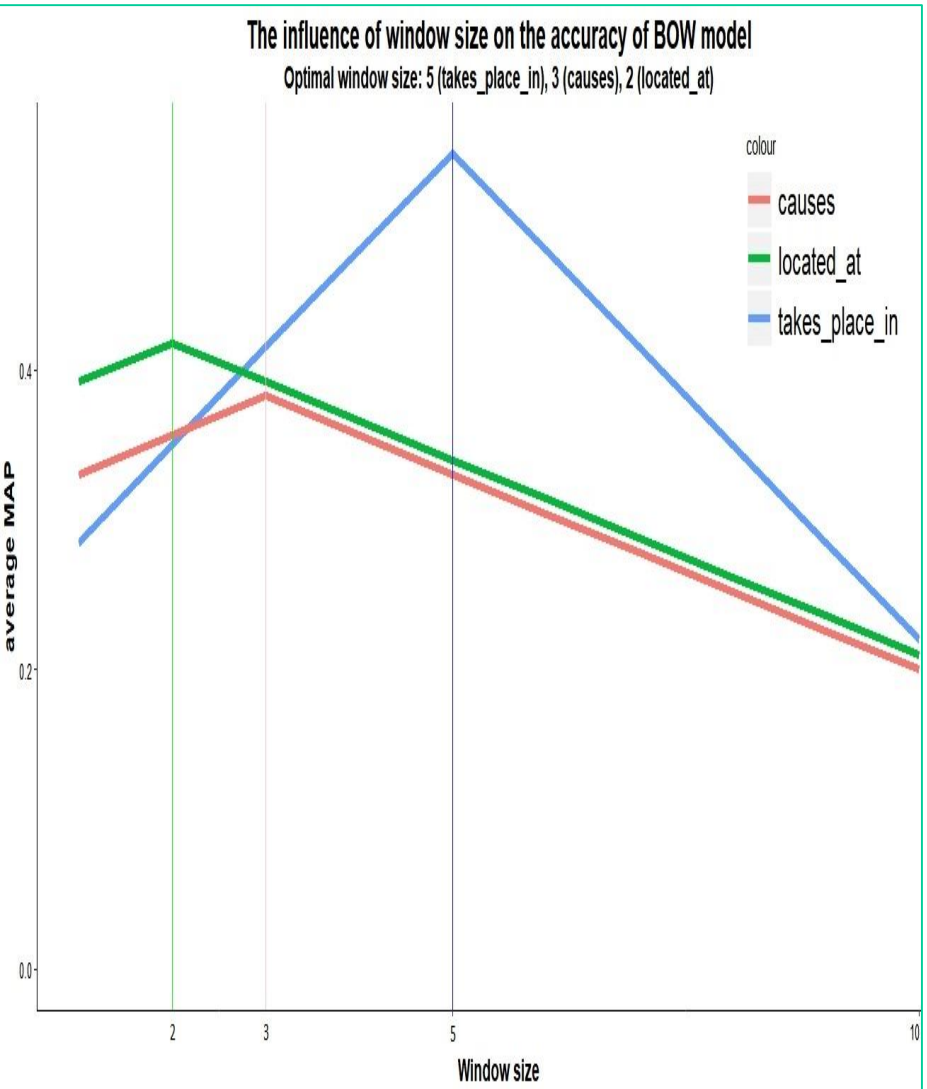
# 4.- Results

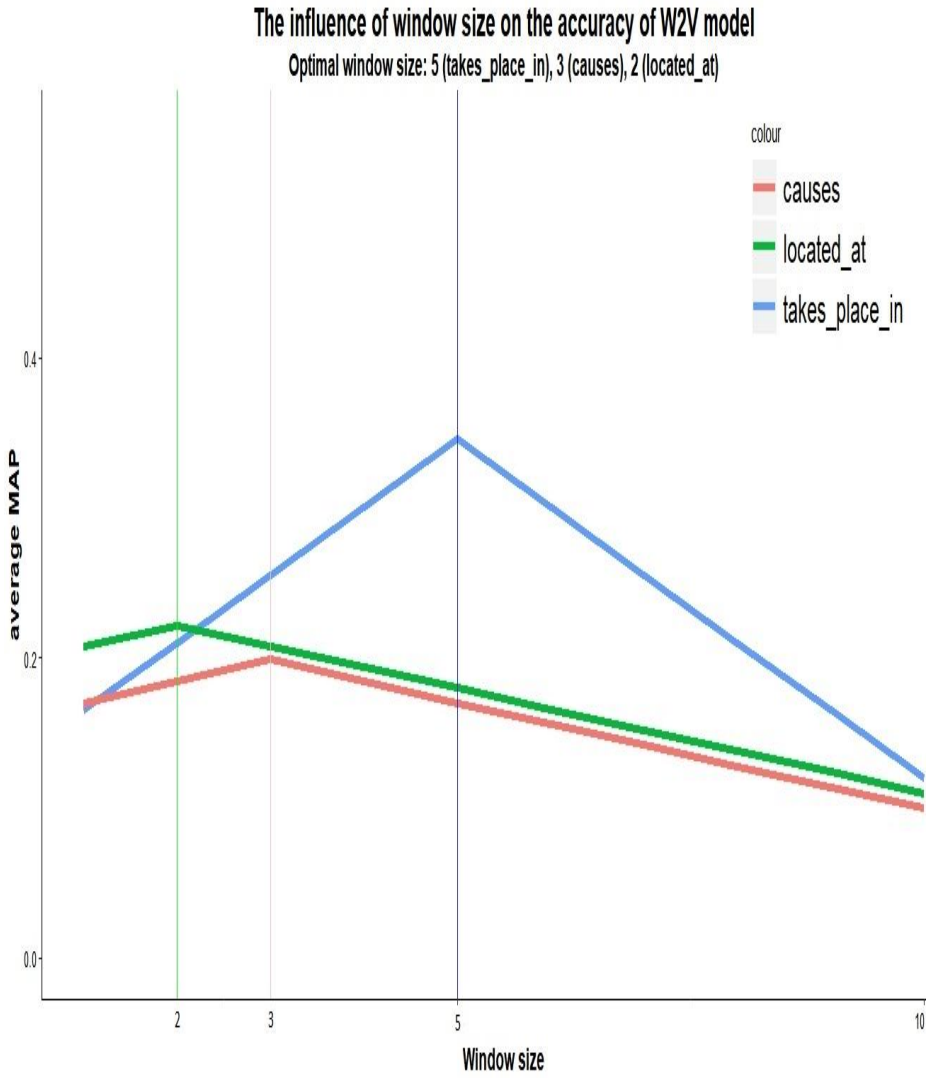
**Comparing count-based and prediction-based models**

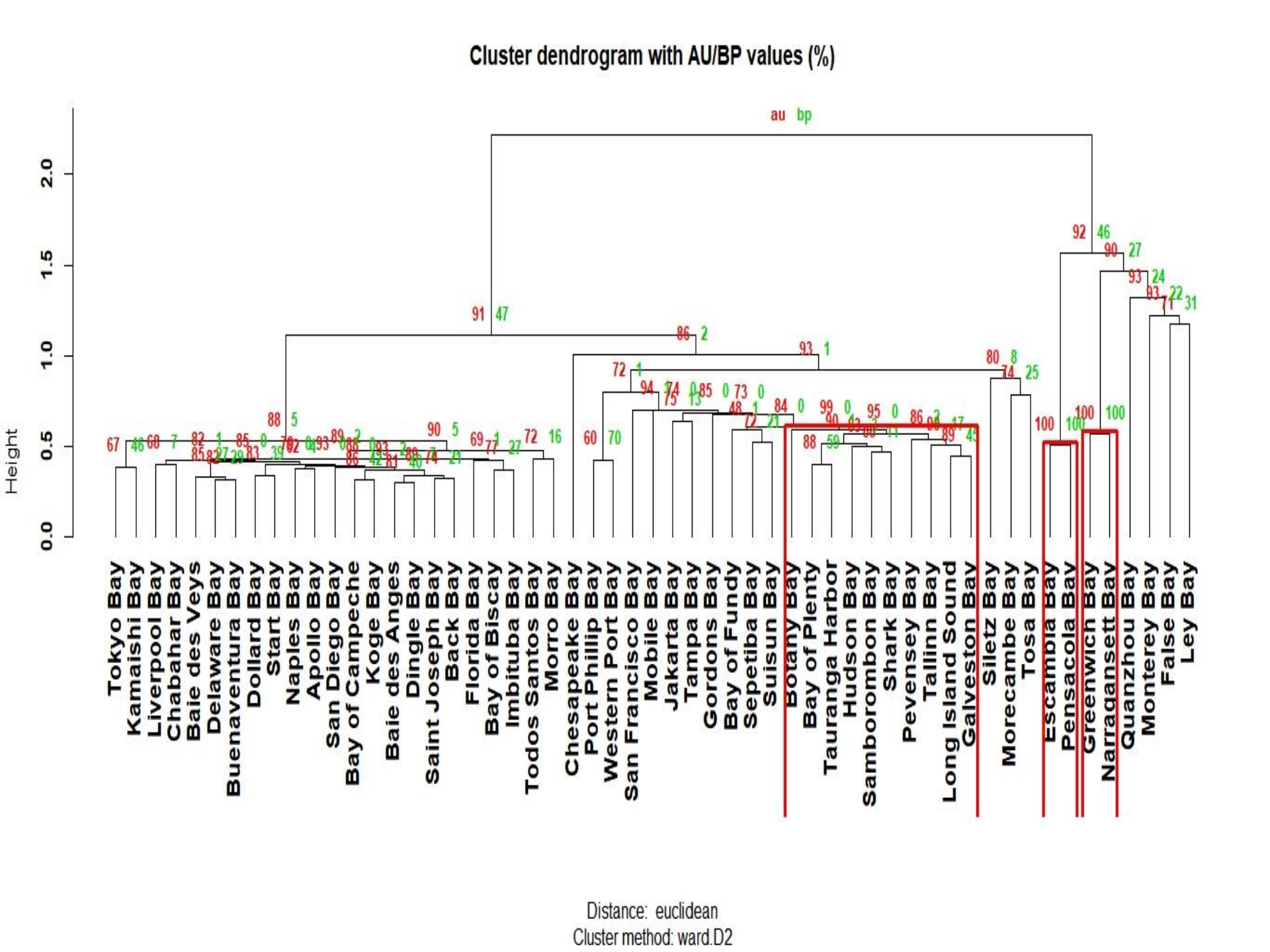
- Now we turn our attention to the influence of the **window size** on the accuracy of both DSMs.

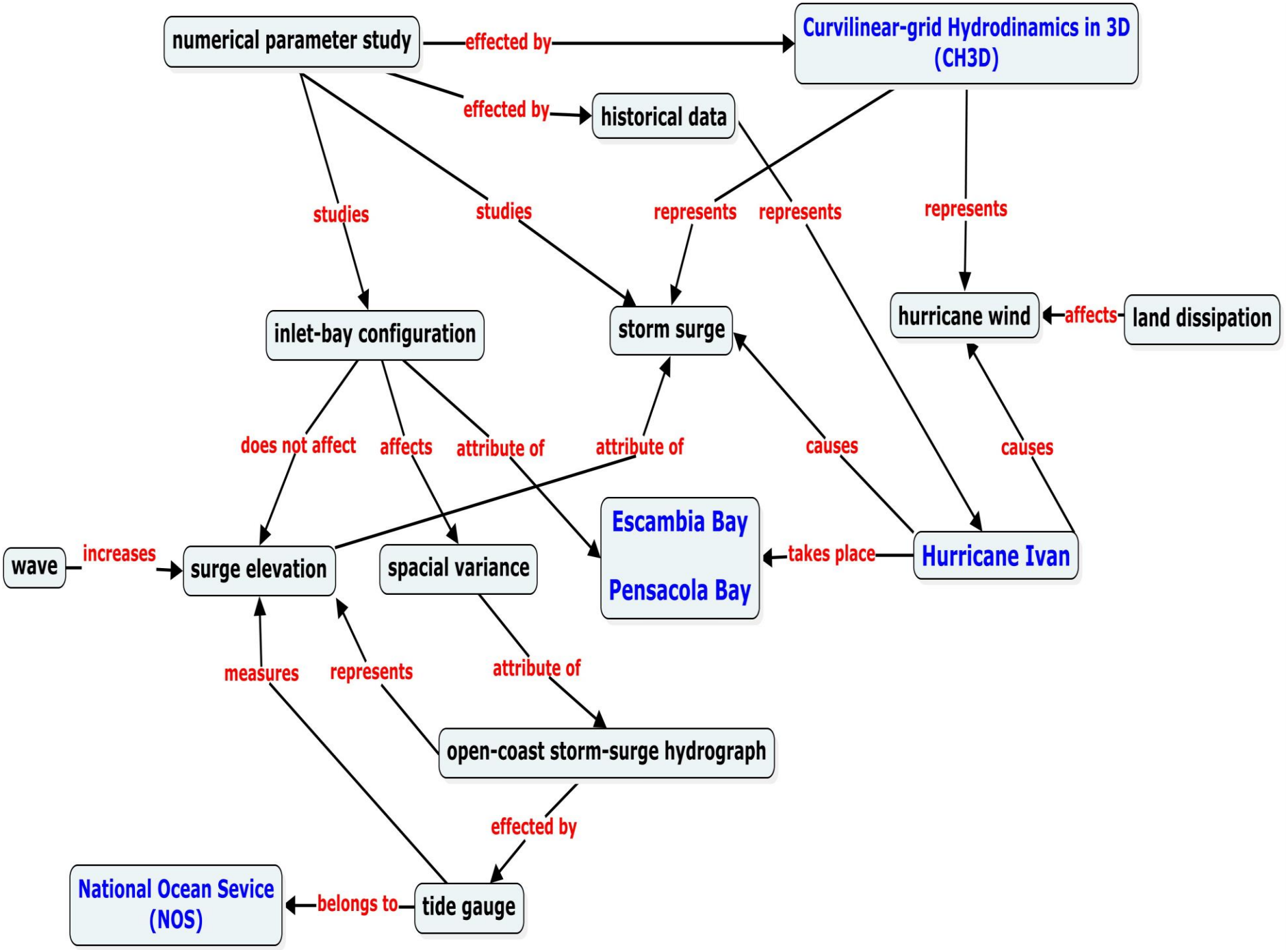
- We use the **average MAP** instead of the maximum in order to determine which settings produce consistently good results.

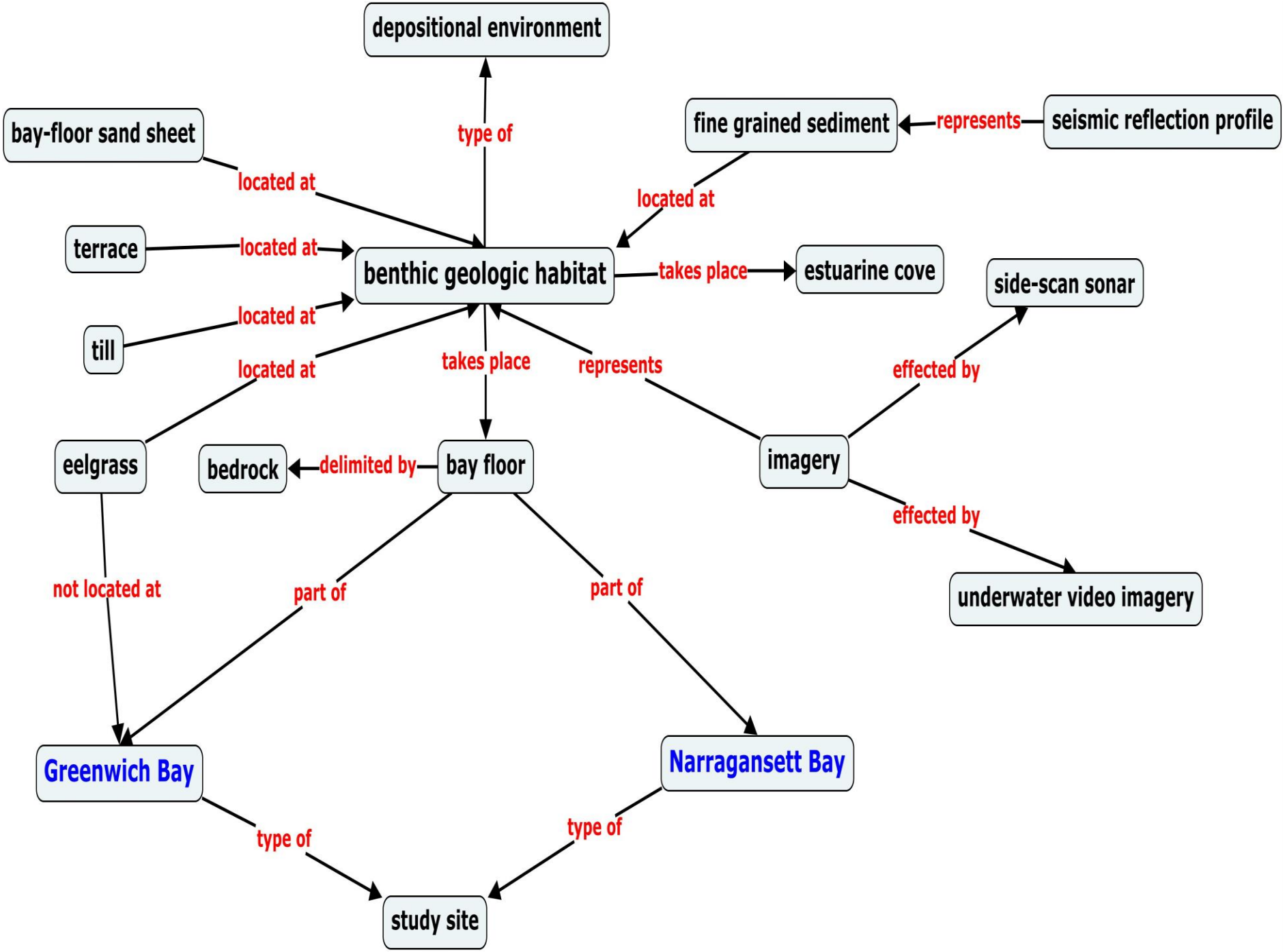
# 4.- Results


The influence of window size on the accuracy of BOW model
Optimal window size: 5 (takes_place_in), 3 (causes), 2 (located_at)
colour
causes
located_at
takes_place_in
average MAP
0.0
0.2
0.4
2
3
5
10
Window size


The influence of window size on the accuracy of W2V model
Optimal window size: 5 (takes_place_in), 3 (causes), 2 (located_at)
colour
causes
located_at
takes_place_in
average MAP
0.0
0.2
0.4
2
3
5
10
Window size


Cluster dendrogram with AU/BP values (%)
au bp
Height
0.0
0.5
1.0
1.5
2.0
Tokyo Bay
Kamaishi Bay
Liverpool Bay
Chabahar Bay
Baie des Veys
Delaware Bay
Buenaventura Bay
Dollard Bay
Start Bay
Naples Bay
Apollo Bay
San Diego Bay
Bay of Campeche
Koge Bay
Baie des Anges
Dingle Bay
Saint Joseph Bay
Back Bay
Florida Bay
Bay of Biscay
Imbituba Bay
Todos Santos Bay
Morro Bay
Chesapeake Bay
Port Phillip Bay
Western Port Bay
San Francisco Bay
Mobile Bay
Jakarta Bay
Tampa Bay
Gordons Bay
Bay of Fundy
Sepetiba Bay
Suisun Bay
Botany Bay
Bay of Plenty
Tauranga Harbor
Hudson Bay
Samborombon Bay
Shark Bay
Pevensey Bay
Tallinn Bay
Long Island Sound
Galveston Bay
Siletz Bay
Morecambe Bay
Tosa Bay
Escambia Bay
Pensacola Bay
Greenwich Bay
Narragansett Bay
Quanzhou Bay
Monterey Bay
False Bay
Ley Bay
Distance: euclidean
Cluster method: ward.D2


numerical parameter study
effected by
Curvilinear-grid Hydrodinamics in 3D (CH3D)
effected by
historical data
studies
studies
represents
represents
represents
inlet-bay configuration
storm surge
hurricane wind
affects
land dissipation
does not affect
affects
attribute of
attribute of
causes
causes
wave
increases
surge elevation
spacial variance
Escambia Bay
Pensacola Bay
takes place
Hurricane Ivan
measures
represents
attribute of
open-coast storm-surge hydrograph
effected by
National Ocean Sevice (NOS)
belongs to
tide gauge


depositional environment
type of
bay-floor sand sheet
located at
fine grained sediment
represents
seismic reflection profile
located at
terrace
located at
benthic geologic habitat
takes place
estuarine cove
side-scan sonar
till
located at
located at
takes place
represents
effected by
eelgrass
bedrock
delimited by
bay floor
imagery
effected by
not located at
part of
part of
underwater video imagery
Greenwich Bay
Narragansett Bay
type of
type of
study site


THANK
YOU

# References


**Baroni, M.; Dinu, G. & Kruszewski, G.** (2014). *Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore (Maryland, USA): ACL, 238-247.

**Benoit, K.** (2018). *quanteda: Quantitative Analysis of Textual Data*. R package version 1.2.0.

**Bernier-Colborne, G. & Drouin, P.** (2016). Evaluation of distributional semantic models: a holistic approach. In: *Proceedings of the 5th International Workshop on Computational Terminology (CompuTerm2016)*, pp. 52-61.

**Bullinaria, J. A. & Levy, J. P.** (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3), 510-526.

**Bullinaria, J. A. & Levy, J. P.** (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods* 44(3), 890-907.

**Evert, S.** (2007). *Corpora and Collocations*. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö (eds.). 2008. *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

**Firth, J. R.** (1957). *A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis* (special volume of the Philological Society), 1952-59: 1–32.

**Flowerdew, L.** (2004). The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings. In U. Connor & T. Upton (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11-33.

**Harris, Z.** (1954). Distributional structure. *Word*, 10 (23): 146–162. (1954).

**Kiela, D. & Clark, S.** (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*. Gothenburg (Sweden): EACL, 21-30.

# References


**Lapesa, G.; Evert, S. & Schulte im Walde, S.** (2014). Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (SEM 2014)*. Dublin: 160-170.

**León-Araúz, P.; Reimerink, A., and Faber, P.** (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem, and P. Fuglewicz (eds.), *Computational Linguistics*. Berlin, Heidelberg: Springer, Studies in Computational Intelligence, 458: 143-161.

**O'Keeffe, A., McCarthy, M. J. & Carter, R. A.** (2007). *From Corpus to Classroom*. Cambridge: Cambridge University Press.

**R Core Team (2017).** *R: A language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. URL: https://www.R-project.org/.