

---

# Summarizing GPS trajectories by salient patterns

Stefanie ANDRAE<sup>1</sup> und Stephan WINTER<sup>2</sup>

<sup>1</sup>Technikum Kaernten, Villach, Austria

<sup>2</sup>Department of Geomatics, The University of Melbourne, Australia

## Abstract

Individuals can easily record their movements in space nowadays with personal navigation devices based on the GPS technology. The recorded data contain in principle detailed information about a traveled route. The information inherent in the data could be used to automatically generate travel diaries. In this paper we study a method to summarize and to symbolize a route according to the cardinal and egocentric directions traveled. The same symbolization method can be applied to other parameters of tracking data, e.g. speed.

## 1 Introduction

For the purpose of human mobile navigation, small, low cost GPS receivers have become available. Affordable prices and reasonable accuracy have increased their use, e.g., in car navigation systems, or for outdoor activities. These devices are able to produce trajectories from quasi-continuous positioning. A trajectory consists of a sequence of time-stamped position and orientation parameters, and represents a travel history of the moving individual that is at least machine readable. Current tracking software can post-process these trajectories. However, tracking software structures and communicates the information about the traveled route in a way that requires some expertise for human analysis and interpretation. Communication in forms close to human cognitive concepts of spatial behavior is still beyond the capabilities of current systems.

Being interested in the *story* in trajectory data, we aim towards an automatic generation of a human-readable travel diary. In principle, a travel diary links geometric knowledge acquired from trajectories with many other resources that provide context, e.g., personal context from a user profile, situational context from the environment, the time of the day, the type of activity, and so on. In this paper we undertake only the first step towards a travel diary: looking at what is in the trajectory data itself. With that approach we assume that results from this research can be combined later with results on deriving context by mining other resources. This assumption holds if trajectory geometry and context are independent, i.e., if context is linked only to patterns in trajectories, but not to the continuous signals. We think that this assumption holds, since the final goal of a diary is verbal explanation, which concerns cognitively salient events, but not the continuum.

Focusing on the geometry of trajectories, we propose a method to produce symbolic (verbal) summaries of trajectory data. The hypothesis is that we can identify salient patterns in trajectory data, and hence, communicate a trajectory by symbols (or words)

characterizing these patterns. Several types of observations can be considered for this classification process, for example the speed of a traveler, or the heading of a traveler, revealing salient patterns in movement. In the following we mainly investigate the heading of a traveler, but the presented and investigated classification methods apply to other observations as well. We will propose a classification method to summarize a trajectory in terms of change in heading. Although these summaries are already useful information in itself, it is to be expected that their value will increase when linked with context data later.

The paper is structured as follows: The next section gives an overview of relevant literature and related work. Section 3 outlines the applied methods and is followed by their implementation in Section 4. The achieved results are presented in Section 5 and discussed in Section 6. We conclude with an outlook in Section 7.

## 2 Literature Review

**Positioning.** To record the movements of a person in space we need to identify his or her location consecutively with a sufficient update frequency. In Table 1 mobile positioning technologies that provide this type of data are outlined (adapted from GIAGLIS et al. 2002). Currently, satellite based positioning technology (GPS) seems to be the most efficient mean to create personal travel histories, at least outdoors, due to small, low cost receivers, sufficient accuracy and global usability.

**Tab. 1:** A taxonomy of mobile location services adapted from GIAGLIS et al. (2002)

Application	Category	Major Technologies
Outdoor	Mobile Telecommunication Network (MTN) dependent	<ul style="list-style-type: none"> <li>• Cell ID</li> <li>• Time of Arrival (TOA)</li> <li>• Observed Time Difference (OTD)</li> </ul>
	MTN independent	<ul style="list-style-type: none"> <li>• satellite based positioning (GPS)</li> </ul>
	Hybrid	<ul style="list-style-type: none"> <li>• Assisted GPS (A-GPS)</li> </ul>
Indoor	Radio based	<ul style="list-style-type: none"> <li>• Wi-Fi (wireless fidelity) networks</li> <li>• Bluetooth</li> <li>• Radio Frequency Identification (RF-ID)</li> </ul>

Satellite based positioning systems use radio signals and can provide an absolute position anywhere on the earth where at least four of the GPS satellites can be clearly observed. The best known and only fully functional system is the US NAVigation Satellite Timing And Ranging Global Positioning System (NAVSTAR GPS). The Russian alternative Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS) with only 12 satellites in operation by March 2004 is proposed to be fully operable in 2007. The intended European Satellite Navigation System GALILEO should be able to offer an operational service from 2008 onwards (EUROPEAN COMMUNITIES 2004).

**Spatio-temporal data.** When adding the time component to spatial data a new data type is created which needs its own data model and algorithms to be stored and analyzed

efficiently. Cheap and ubiquitous data transfer will enable the creation of large archives of personal spatio-temporal data. The indexing and querying of spatio-temporal information from moving objects is addressed, for example, in the work of PFOSE et al. (2000). MILLER (2005) extends this and similar approaches by defining rigorous time geographic concepts and relationships based on the conceptual framework of time geography (HÄGERSTRAND 1970). MILLER'S definitions form a measurement theory, able to provide a basis for spatio-temporal queries in future location-based services.

**Visualizing and analyzing spatio-temporal data.** GIS software that is able to perform more complex operations on spatio-temporal data is relatively rare on the GIS market (DODKINS 2004). Software to visualize GPS trajectories comes traditionally with the GPS devices and enables users to display their traveled trajectories statically or dynamically on a map, or to create trajectory profiles. In her study, DODKINS evaluates the effectiveness of the ESRI ArcGIS Tracking Analyst software for analyzing and visualizing the spatial behavior of a human traveler. DODKINS evaluates the software as satisfactory to carry out basic visualization and animation, like dynamic animations, static data clocks and histograms. She also discusses slow processing time when working with large datasets, and a lack of comprehensive statistical analysis. However, through visualization it is possible to communicate human travel behavior visually. For instance, it becomes visible that there are locations where the individual was staying for a longer period of time, or there are route segments traveled with different speeds. Some of these events or activities happen recursively at similar day times or other temporal intervals, which demonstrates that there are individual behavioral patterns implied in the data. MOUNTAIN and DYKES (2002) agree that visualization is a useful tool to reveal structure. In their work they elaborated on visualization and analysis of GPS trajectories and developed a series of methods to extract and represent information visually.

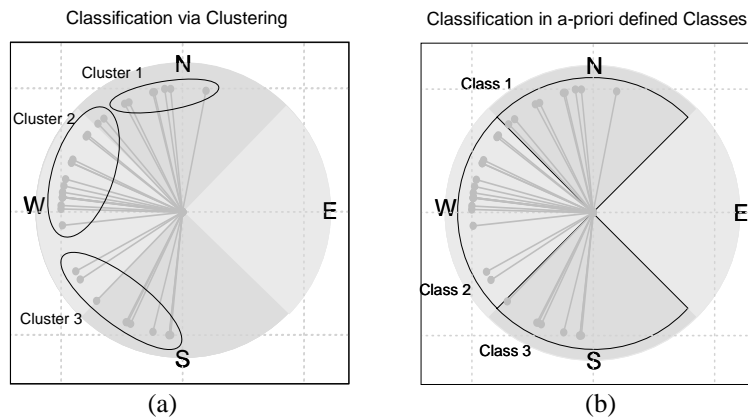
**Clustering of spatio-temporal data.** Clustering is an unsupervised classification method for grouping similar objects. Basic knowledge about the technical background, different clustering methods and their diverse fields of applications can be gained from text books (e.g., HARTIGAN 1975; GORDON 1999; EVERITT et al. 2001). A symbolization method using clustering techniques has already been proposed by JIN (2004) to automatically identify typical sub-trajectories in a mobile agent trajectory. JIN'S method segments a trajectory into sub-trajectories, clustering them for their similarity according to their proximity measured in Euclidian distance. As a result the original trajectory can be symbolized by replacing the original sub-trajectories with their respective cluster centers, which are the mean of all the elements grouped into one cluster. Applying the method on directions, JIN notes that a rotation needs to be applied on the extracted segments, so that the symbols derived represent rather egocentric movements (such as "left", "straight", "right").

**Direction concepts.** A qualitative cardinal representation of traveled directions is reasonable for instructions in large-scale geographic space (FRANK 1996). For trajectories collected in structured space, like cities, where the individuals follow predefined paths, egocentric direction symbols are proposed to be more valuable. KLIPPEL et al. (2004) investigated the mental direction concepts of humans in city street networks. With an experimental method they revealed the number and size of sectors in which humans distinguish directions.

### 3 Approach

To identify and describe sequentially the cardinal or egocentric directions of a traveler, there are several options. All are based on the information at an observed tracking point.

Regarding the classification method, we distinguish clustering techniques, which adapt to given trajectory data, and classifications in a-priori defined classes, which do not adapt (Figure 1). In this paper we compare adaptive clustering methods, and non-adaptive classification. We expect from adaptive methods more sensitivity for patterns, and from non-adaptive methods results according to a general ontology.



**Fig. 1:** Two different classification methods: (a) The three clusters found with a agglomerative hierarchical clustering method. (b) Classification into a-priori defined classes.

JIN proposes distance based clustering methods, such as  $k$ -means or hierarchical clustering. The  $k$ -means method can start with randomly chosen initial centers. After a maximum number of iterations, switching objects from one cluster to another in order to minimize the sum of squares in all cluster centers, the algorithm terminates with a grouping that is locally optimal. A drawback of this quick and commonly used method is that it assumes that the user specifies the number of clusters  $k$  in advance. Furthermore it depends on the arbitrary choice of initial cluster centers, so that a different initial choice might lead to another final classification result (HARTIGAN 1975).

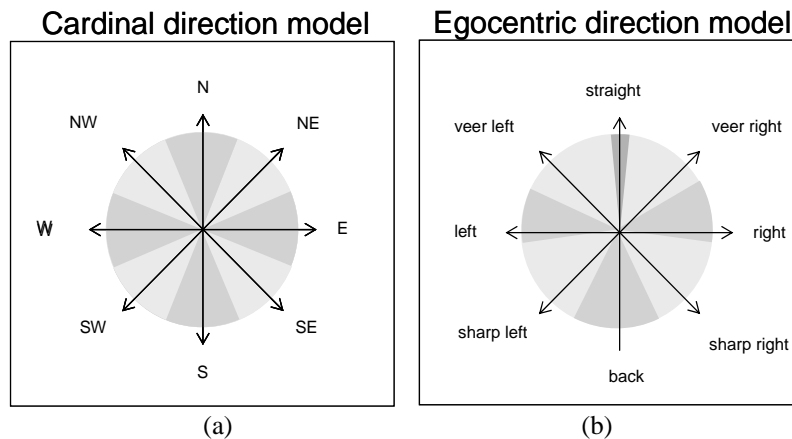
Applying hierarchical agglomerative clustering with Euclidian distance as proximity measure is still adapting to the data, but independent to any start values. The algorithm starts with an initial partition of each element into a single cluster, at each stage the most similar elements or clusters are fused, and it finishes with one cluster containing all the elements. This method avoids the dependency on an arbitrary initial choice. However, the crucial question of the appropriate number of clusters  $k$  remains. Although there exist formal methods to determine the optimal number of clusters  $k$ , like *stopping rules* for hierarchical methods or *evaluation graphs*, like the  $L$  methods (SALVADOR and CHAN

2003), we chose an informal method which involves studying the graphical representations of data and cluster centers (Gordon 1999).

In order to cluster directions we define the variables to be clustered as the Cartesian coordinates  $x = \cos(t)$  and  $y = \sin(t)$ , rather than the angle  $t$ , which is the change in direction between two consecutive data points in a trajectory. These observations, measured in degrees, are data of a cyclic scale of measurement, which have to be treated different than data of an interval or ratio scale of measurement (CHRISMAN 2002). In this case the similarity measure needs to be modified (MARDIA 1972). However, a cyclic representation of direction ( $x = \cos(t)$ ,  $y = \sin(t)$ ) enables use of an Euclidian distance based clustering algorithm. In order to cluster egocentric directions traveled, the relative change in direction to the previous heading was calculated.

In comparison to the above, a non-adaptive classification is fixed in the number of classes and in the definition of classes. Classes are defined by an agreement in the information community; with respect to directions, for example the cone shaped direction model for the cardinal directions north, east, south and west (FRANK 1996), or an egocentric direction model, such as the conceptual model of KLIPPEL et al. (2004) for the egocentric directions straight left, right, veer left/right, sharp left/right. In these cases the groups are fixed into four and eight classes respectively (see Figure 2). The cardinal model in Figure 2 (a), divides the unit circle into eight areas (the four cardinal directions plus north-east, north-west south-east, south-west). The classes for egocentric directions are derived from the direction model of KLIPPEL et al. (2004), see Figure 2 (b). Their conceptual model was designed based on the results of a grouping task paradigm on discrete direction icons. Therefore their model exhibits gaps of some degrees between each identified direction. As no further information is available we assume the bisecting line of those gaps as the border between classes. Additionally we named the egocentric direction symbols according to their model, although they note that the proper naming of the direction sectors will need to be addressed in future work.

After these variables are grouped into clusters, or into a-priori defined classes, every element is substituted by a prototypical symbol representing each class. In this case, clustering the mean of all elements in each cluster is used as the representative symbol. To name the symbols we use the cardinal or egocentric direction model introduced above (Figure 2).



**Fig. 2:** A-priori defined classes for directions: (a) A cardinal direction model of eight cones, and (b) the direction model taken from KLIPPEL et al. (2004).

In the case of a-priori defined classes, the a-priori defined direction axes can be used. A sequence of the same symbols is then summarized to derive a qualitative (verbal) description of the route. Depending on the type of direction – cardinal or egocentric –, different verbal and visual descriptions and of the trajectory can be achieved:

- **Cardinal directions.** A summary of the main cardinal directions traveled can be derived by regarding consecutive sequences of the same symbol as one segment. A verbal description is derived by calculating the distance between consecutive points of the same symbol and naming the direction according to the symbol, for example “1.804 m north-west, 982 m north...”. A visual representation is achieved by assigning a different color to each symbol and coloring the data points and the connecting line to their successor in that color. Additional information like minimum, maximal, average speed traveled in each segment is available in the trajectory data.
- **Egocentric directions.** For trajectories collected in structured space, such as cities, egocentric direction symbols on decision points are proposed to be more valuable in terms of communication to people. Consecutive sequences of the class “straight” are summed up in the same way as cardinal directions. All other classes (“veer right/left”, “sharp right/left”, “right/left”) are supposed to represent outstanding turns and so do not occur in a sequence of symbols. They were mentioned individually when occurring in sequence, for example “1.702 m straight, veer right, veer right, 58m straight, ...”. For a visual representation each data point is plotted in the particular color, size or shape assigned to its class.

Note that the naming of the classes derived via clustering, the same models were used for the a-priori classification. It can happen that several classes fall into one cone of the direction model and thus can be labeled equally.

## 4 Implementation

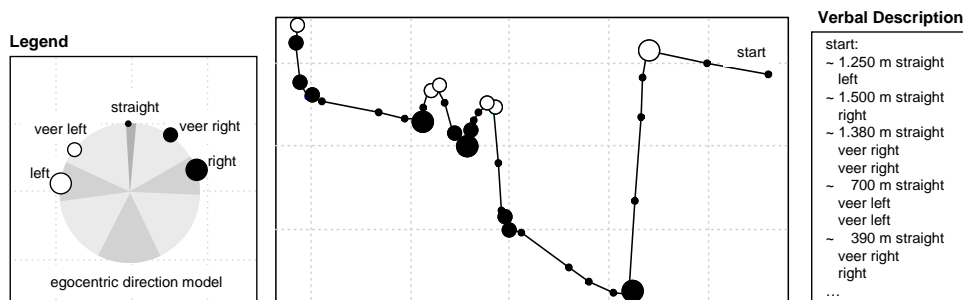
The open source software *R* was used as an environment for the statistical computation (THE R FOUNDATION 2005). *R* was chosen as it is free software under the GNU license and provides powerful statistical and graphical functions. The additional package *shapefiles* by STABLER (2003) was used to import GPS trajectories. The *k*-means algorithm of HARTIGAN and WONG (1979) and the agglomerative hierarchical method, both already implemented as the functions *kmeans()* and *hclust()*, in the *stats* package of *R* were used. Further functions were implemented to process the trajectory data to derive variables for cardinal and egocentric directions, as well as functions to visualize the results.

The functions were applied on two data sets. The first consisted of 3.550 map matched data points, which represent 250 km of trajectories, collected by a car navigation system over five days. The second consisted of 7.035 data points, which represent more than 950 km of trajectories, collected with a portable navigation device over ten days. Both data sets were collected in the region of Melbourne, Australia and represent the typical output of commercially available navigation systems.

## 5 Results

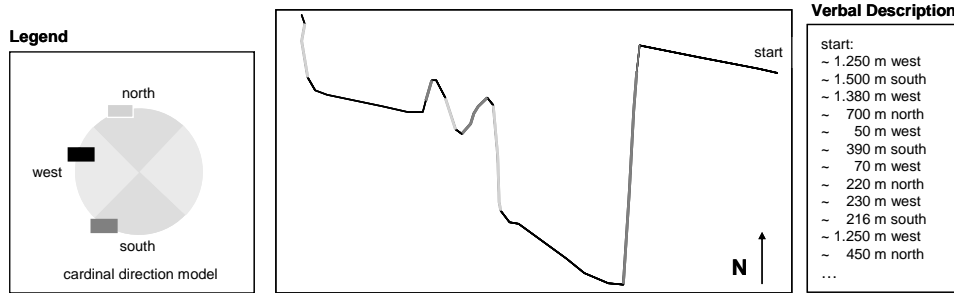
The clustering algorithms grouped the data points of the GPS trajectories into classes of cardinal and egocentric classes. Figures 3 and 4 show the results of applying the adaptive method on a section of the trajectory collected with the personal navigation device. To reduce the noise around stops a distanced based filter is applied which eliminates all data points which did not fulfill a maximal distance of 500 m, a minimum distance of 50 m, a maximum deviation of 10 m or a maximum angle of  $25^\circ$  to their successor. The hierarchical agglomerative clustering, using *average* linkage as the inter-cluster distance measure, groups the data points into the five (egocentric) and three (cardinal) classes respectively shown in the legend.

### Visualization and description of egocentric direction pattern of a GPS trajectory



**Fig. 3:** A trajectory classified into egocentric direction classes computed via clustering.

**Visualization and description of cardinal direction pattern**

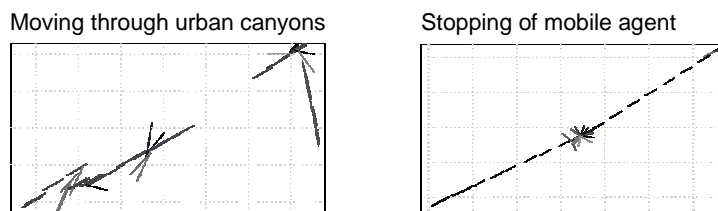


**Fig. 4:** A trajectory classified into cardinal direction classes computed via clustering.

The legend shows the location of the cluster centers in the direction model in the color and size used in the visualization in the middle. The center is labeled according to the direction model cone it falls into. The verbal description on the right hand side states the distance between consecutive points of the same class followed by the class label.

## 6 Discussion

With the described classification methods it is possible to derive verbal descriptions for trajectories. Salient patterns, related to the direction of the movement, are revealed in the sequence of classified data points (e.g., [NNNEEE] shows a change in direction). Considering the data, GPS trajectories include very detailed information about movement on earth's surface, but at the same time inevitable noise. Accumulation of noise happens in particular when a moving agent performs a stop or moves through areas where the signal is obstructed. During these stops, the positions recorded are scattered around the actual location, because of the noise inherent in the GPS technology. In areas with poor signal reception, positions recorded are often far away from the actual position or are missing altogether (Figure 5).



**Fig. 5:** GPS data problems visualized by cardinal direction symbols.

However, this problem can be overcome by existing techniques for digital signal processing, such as filtering, or techniques for intelligent navigation, such as map matching or the use of additional inertial sensors (e.g. SCOTT-YOUNG and KEALY 2002). For the purpose of directions traveled, a technique should reduce noise, identify and ignore stops, but should preserve important turns.

Furthermore clustering results depend on the distance method and number of cluster centers  $k$  chosen. We experienced that for hierarchical clustering, with its three most common inter-group distance measures *single*, *complete* and *average* linkage, *average* linkage can be recommended, since it yields to more evenly distributed classes. An appropriate choice of  $k$  depends on each individual trajectory, and no specific value  $k$  can be recommended for general use.

Clustering methods, which adapt on the data, have advantages in summarizing trajectories. Consider Figure 3: One cluster center is classified *straight*, so all elements in this cluster are labeled *straight*, although, classified via the non adaptive method, they may belong to other classes of the a-priori given egocentric classification model. This means that the hierarchical clustering, by adapting to the current data set, combines more segments into *straight* than the classification into the a-priori defined class. A classification by the a-priori defined class is more sensitive, since this class is particularly small. It yields a longer verbal description, due to more points classified as *veer right/left*. However, in our case we found the distinctions hypersensitive from a cognitive point of view.

## 7 Conclusion

In this paper we focused on qualitative information which can be extracted simply out of the geometry of mobile agent trajectories. We applied two methods to produce symbolic (verbal) summaries of trajectory data in terms of changes in directions. We used clustering methods which adapt the data and a-priori defined classes to classify GPS trajectories. The clustering method lets us identify salient patterns in trajectory data, enabling communication of these patterns by symbols (or words) and can be applied on other observations such as speed.

## References

- Chrisman, N., 2002: Exploring Geographic Information Systems. John Wiley & Sons, New York.
- Dodkins, J., 2004: Exploring the movement of a human traveller from continuous spatio-temporal data using ESRI ArcGIS Tracking Analyst software. Technical report, Department of Gematics, The Universtiy of Melbourne, Melbourne, Australia.
- European Communities, 2004: GALILEO - Programme phases, [http://europa.eu.int/comm/dgs/energy\\_transport/galileo/programme/phases\\_en.htm](http://europa.eu.int/comm/dgs/energy_transport/galileo/programme/phases_en.htm). Last accessed 11. January 2005.
- Everitt, B.S.; Landau, S.; Leese, M., 2001: Cluster Analysis. Arnold, London.

- Frank, A.U., 1996: Qualitative Spatial Reasoning: Cardinal Directions as an Example. *International Journal of Geographical Information Systems*, 10 (3): 269-290.
- Giaglis, G.M.; Pateli, A.; Fouskas, K.; Kourouthanassis, P.; Tsamakos, A., 2002: On the Potential Use of Mobile Positioning Technologies in Indoor Environments, In *proc. of the 15th Bled Electronic Commerce Conference*, Bled, Slovenia.
- Gordon, A.D., 1999: Classification. *Monographs on Statistics and Applied Probability*, 82. CHAPMAN & HALL/CRC, New York.
- Hägerstrand, T., 1970: What about people in regional science?, *Papers of the Regional Science Association*, pp. 7-21.
- Hartigan, J.A., 1975: *Cluster Algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.
- Hartigan, J.A.; Wong, M.A., 1979: A K-Means Clustering Algorithm. *Applied Statistics*, 28 (1): 100-108.
- Jin, X., 2004: Symbolization of Mobile Object Trajectories with the Support of Motion Data Mining, In *proc. of the 23rd International Conference on Conceptual Modeling (ER2004)*, Shanghai, China.
- Klippel, A.; Dewey, C.; Knauff, M.; Richter, K.-F.; Montello, D.R.; Freksa, C.; Loeliger, E.-A., 2004: Direction Concepts in Wayfinding Assistance Systems, In *proc. of UbiComp 2004: 6th International Conference on Ubiquitous Computing*, Nottingham, UK, pp. 1-8.
- Mardia, K.V. (Ed.), 1972: *Probability and Mathematical Statistics - A Series of Monographs and Textbooks*. Statistics of Directional Data. ACADEMIC PRESS INC. LTD., London.
- Miller, H.J., 2005: A measurement theory for time geography. *Geographical Analysis*, in press.
- Mountain, D.; Dykes, J., 2002: What I Did On My Vacation: Spatio-Temporal Log Analysis With Interactive Graphics And Morphometric Surface Derivatives (Abstract), *GIS Research UK 10th Annual Conference*, Sheffield, UK.
- Pfoser, D.; Jensen, C.S.; Theodoridis, Y., 2000: Novel Approaches to the Indexing of Moving Object Trajectories, In *proc. of the 26th International Conference on Very Large Databases*, Cairo, Egypt, pp. 395-406.
- Salvador, S.; Chan, P., 2003: Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. Technical Report CS-2003-18, Dept. of Computer Sciences, Florida Institute of Technology, Melbourne, Florida.
- Scott-Young, S.; Kealy, A., 2002: An Intelligent Navigation Solution for Land Mobile Location Based Services. *The Journal of Navigation*, 55: 225-240.
- Stabler, B., 2003: R package: shapefiles, <http://mirror.aarnet.edu.au/pub/CRAN/>. Last accessed 22. December 2004.
- The R Foundation for Statistical Computing, 2005: The R Project for Statistical Computing, <http://www.r-project.org/>. Last accessed 11. January 2005.