The Journey is More Important Than the Destination: Finding Scenic Routes Using TomTom Routing Data

Jayme Bird
Vrije Universiteit Amsterdam
De Boelelaan 1105
Amsterdam, the Netherlands
j.bird@student.vu.nl

Stephen Nicholas Swatman
Vrije Universiteit Amsterdam
De Boelelaan 1105
Amsterdam, the Netherlands
s.n.swatman@student.vu.nl

ABSTRACT

Be it for business or for pleasure, road travel constitutes a significant part of the lives of many in modern society. The environment in which we drive can have a large impact on mental health and happiness. Providing a method of finding scenic routes between two points may help people identify ways to increase the enjoyment of their commute and to decrease their levels of stress. While previous research has attempted to provide such methods, all have relied on metadata and volunteered geographic information (VGI).

In this paper, we describe our attempt to automatically classify scenic and non-scenic routes based on LIDAR point cloud data using VGI only as a form of ground truth. While we achieve limited success using LIDAR data we believe we succeed in automatically identifying scenic routes through nature using a method with a higher reliance on VGI. We extract features using enriched core base maps to obtain categorized scenic segments. We match and filter these using aggregated density points of 1.6 million planned itineraries making up billions of coordinate pairs. The result is the most popular scenic segments in Europe that people have planned routes to. in Addition, we use these scored routes to perform several additional experiments including point of interest (POI) ranking.

Keywords

Apache Spark, Scenic route, Driving enjoyment, Classification, GIS, Geographic information

1. INTRODUCTION

It is a common saying that the journey can be as important as the destination when it comes to travel. Even in our day to day lives the journeys that we undertake have an important impact on our wellbeing. For example, an unpleasant daily commute can lead to a decrease in mental health while people who enjoy their commute report increases in overall happiness [8, 7]. Two of the factors which most strongly influence the enjoyment of driving are the scenery of the drive and the freedom experienced by the driver [9].

One explanation for the fact that a third of drivers report being dissatisfied with their commutes, then, is that the routes that they choose to drive on are not scenic. Why is this? It is possible that drivers choose monotonous routes over more scenic alternatives but it may also be the case that drivers are simply not aware of scenic routes to their destinations. Indeed, resources on which routes are scenic and which are not may not be available for all regions and such information may be hard to discover through methods other than word of mouth and time-consuming trial and error.

In this paper, we attempt to find a method to compute scenic stretches of road using the Apache Spark big data framework. To this end we use data from *TomTom*, a Dutch digital mapping company and manufacturer of car navigation systems. We combine this data of routes planned using the *TomTom MyDrive* route planning system with a dataset extracted aesthetic routes from the TomTom basemaps. Additionally we perform an experiment of scenic route detection using point cloud elevation models.

Our goal is to create a visualisation of the popular aesthetic tourist routes in Europe split into several categories including routes of natural, cultural, national and regional significance. By providing a map of scenic road segments we aim to make it easier to find and plan trips along the most scenic stretches of the European road system. We believe such a visualisation can also be useful to both tourists and long distance commuters. Ultimately, allowing people to choose to drive on more scenic routes may prove a boon to their happiness and mental health regardless of their motivations.

To gain insight into existing methods we discuss previous literature on both classification methods and relevant big data technologies in Section 2. We expand and detail our research questions in Section 3 and we describe the data sets used in Section 4. The methods and programs developed to process and visualize these datasets are described in Section 5. We briefly discuss experiments done in Section 6 and provide concluding remarks in Section 7.

2. RELATED WORK

While substantial effort has gone into researching the effects of unpleasant commutes [8, 7, 9] there is relatively little published research on what constitutes a pleasant driving environment. In this section we address factors that determine the scenicness of a route and attempt to classify

previous methods of finding scenic routes according to various different levels of human effort. Additionally, we briefly touch upon technical methods that can be used to process data at large scales, both generally and specifically for geographical data.

Alivand, Hochmair, and Srinivasan [1] present a model for the scenicness of routes and find that accurate classification of scenic routes can be achieved according to a small number of variables; the presence of nearby bodies of water, mountains, parks and the level of urban development in the area. These variables are determined to be the most significant out of a longer list of variables of which some are highly niche (such as the presence of places of worship). While a small sample size of 96 routes, all of which in California, does not confirm that these variables are generally usable, they provide a starting point for new models in all but the most extreme regions of the world.

A method of finding scenic routes through cities is given by Quercia, Schifanella, and Aiello [3]. The method used in their research involves crowdsourced classification of images of streets in cities. Such images are presented to volunteers who then score the scenicness of the images. The scenicness of road segments is then represented in a graph format such that a scenic route can be found between any two points in the city. The authors address an important shortcoming of their research, the fact that images must me manually classified, by also considering metadata on the popular photo sharing site *Flickr*. The sentiment of comments and the number of *favourites* are combined with geotags to automatically compute the scenicness of locations. While the authors achieve accurate results using this method there is still a strong reliance on user-entered data in the form of photographs. We conclude that such a method can work well for cities but may fall short for less frequently visited or photographed areas.

A similarly photograph-based approach is presented by Runge et al. [10]. Here, images are automatically classified using deep learning techniques to identify six different features: mountains, fields, bodies of water, nature, sightseeing locations and non-scenic features. The proposed *Autobahn* system retrieves images from the *Google Street View* service which provides panorama views of roads all around the world. While human effort is required to create and update the panorama images, there is no reliance on human classification of scenicness through comments or *favourites* as there is in the research of Quercia, Schifanella, and Aiello [3].

The previously mentioned research by Alivand, Hochmair, and Srinivasan [1] does not rely on photography and instead classifies the scenicness of routes using points of interest and polygons extracted from various datasets. While some such datasets may be created partially automatically they require human moderation to ensure accuracy. We are not aware of any scenic route classifications methods that are fully automatic and that do not depend on volunteered geographic information (VGI) or metadata. In an attempt to develop such a method, we will now also briefly explore methods for big data engineering in general and for geographical data specifically.

Apache Spark is a powerful tool for analysis of very large data sets [11]. Spark uses distributed data structures of various levels of rigidity to provide high and scalable performance. Specifically, resilient distributed datasets (RDDs), DataFrames and simple DataSets are read-only data structures that can be operated on, in memory, by the machine that also stores the data. This minimizes the need for data transfer between machines in a cluster and can potentially lead to much higher performance compared to similar technologies such as Apache Hadoop.

Geospark [4] is a Spark library for the processing of geospatial information. Geospark features support for different geospatial datatypes such as points, lines and polygons and can store RDD's of such types. However, it must be noted that the number of supported input file formats is limited and does not include the LAS format which is commonly used for point cloud data. Geospark has support for a wide variety of spatial operations and partitioning methods including distance-based join queries for Spark SQL and quadtrees.

A different Spark library specifically designed to support point cloud data is *IQmulus* [12]. While IQmulus relies on outdated versions of the Spark framework at the time of writing it does have support for LAS, XYZ and PLY file formats. An important weakness of IQmulus is the lack of documentation that is available. Additionally, there is no explicit support for very large point clouds which are split into multiple files. Boehm, Liu, and Alis [5] detail a procedure which can be used to load such data sets into Spark in their entirety. Similarly, Liu and Boehm [6] show a method of loading split point clouds into a Spark pipeline specifically for classification purposes.

3. RESEARCH QUESTIONS

The goal of this research is to identify scenic routes, and determine the most popular scenic routes by analysing a large set of saved planned routes from TomTom MyDrive, a route planning service.

1. Which of the routes stored in our data are the scenic/tourist routes?
   (a) Can we automatically classify scenic routes using LiDAR point cloud data?
2. What are the most popular routes planned within the data set.
3. What are the most popular scenic routes within the data?

In order to answer the first research question, we need to identify the scenic routes within the 1.6 million saved routes that make up the data set. First, we need to determine what constitutes a scenic/tourist route, and then determine which roads are routes are scenic. As a sub-question, we wish to determine if this can be automatically classified using LiDAR point cloud data. Once the first question is answered, we need to determine what the most popular routes are within the dataset, and finally apply this popularity metric to filter the scenic routes to find the most popular scenic routes within the data set.

In order to address the research questions, specific software, tools and data sources are required. For this project, we will be using Spark, TomTom MyDrive routes, AHN LiDAR Point Clouds as well as TomTom base map information. The results will be presented in this paper, and within a web based visualization.

4. DATA

This research is conducted on three primary data sets. The first is TomTom MyDrive data, which consists of 1.6 million saved routes that users have planned using the MyDrive Route Planning website. The second major dataset used is locations of scenic/tourist road features extracted from Tom Tom Base Maps. The third set of data is the Actueel Hoogtebestand Nederland (AHN2/AHN3) LiDAR point cloud data of The Netherlands.

4.1 TomTom MyDrive

The main set of data is the TomTom MyDrive data, which spans 2013-2016 and consists of 1.6 million saved routes, with roughly 1.5 billion coordinate path points and is about 150GB in size. The data is stored in MongoDB, and compressed using the WiredTiger storage engine. The coverage of the data is primarily within Europe, making up 1.4 million of the total 1.6 million saved routes. Due to this, our analysis will focus on the European routes.

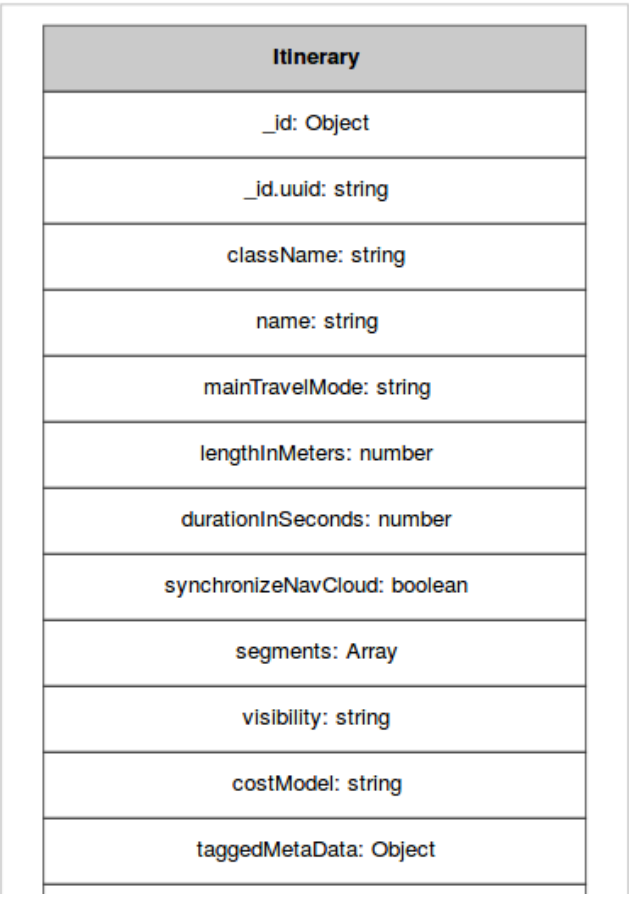


Figure 1: Schema Analysis of TomTom Route Itinerary Data

Route distance and time features such as LengthinMetres and DurationinSeconds(integer);

synchronizeNavCloud Identifies if the route was sent to a TomTom GPS device or just saved as a planned route in MyDrive

Name features such as Name of Route (String); Route ID (String):

Cost Model Cost model for route such as Fastest, thrilling or exact;

Visibility Boolean value of public or private. Currently public routes are only from the route library of 100 scenic routes.

Segments Array of coordinates. Segments makes up the collection of longitude and latitude coordinates embedded as path points and waypoints. A single saved route can have several hundred coordinate pairs.

4.2 TomTom Base Map Extracted Layers

TomTom Base Maps provide a valuable source of detailed map information which are enriched by map makers who drive around and note specific names or features associated with a stretch of road. Within the context of discovering scenic routes planned within the TomTom MyDrive data, the base map data provides a good source of information. The map data was obtained from a compressed vector tile format, and was in total 28GB and was exported as layers from the map. The scheme of the map data is shown below.

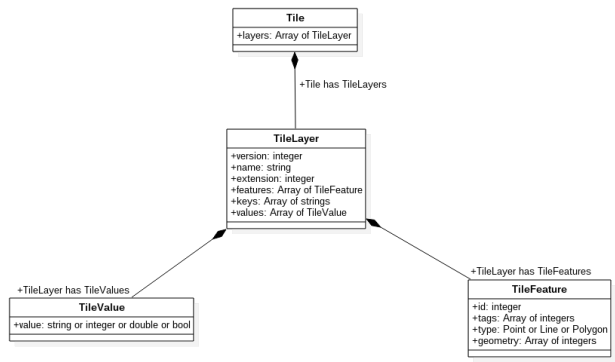


Figure 2: Schema of TomTom Base Maps Data

The features and name attributes are of particular use for our research as it contains detailed map data of landmarks, points of interest, area and road information as well as cartographic Labels, if it's of cultural, or regional importance and details of tourist roads and nature that will be useful in determining where scenic/aesthetic stretches of road may exist.

4.3 GPSies

An additional sub data set of the TomTom MyDrive data was discovered during exploratory analysis. The data originates from GPSies.com, a website built for route recommendations. 7000 routes were discovered within the TomTom MyDrive data as users upload routes as GPX files to MyDrive in order to send the route to their PND (personal Navigation Device). The routes were discovered due to the fact that when a user exports a route from the GPSies.com website as a GPX file, it automatically adds the word gpsies.com at the end of the route name. This data was, therefore, easily extracted by querying routes that contain 'gpsies.com'.

The routes from GPSies.com are routes that are highly ranked and recommended as being particularly enjoyable or scenic. They feature a star based ranking system. While this ranking is lost when the route is saved into MyDrive,

we aim to create a new ranking system through clustering and aggregation to find the top routes from the GPSies.com subset.

4.4 LIDAR Data

linear vegetation elements such as tree lines and hedges in rural areas across Europe could be associated with roads that are considered to be scenic, or aesthetic. The rationale for utilizing this data while searching for aesthetic or scenic routes is that these features are considered being a part of the cultural landscape and function as corridors into roads less traveled that that are more enjoyable; the goal of this project and analysis. An inventory of the spatial distribution of these features and elements can, therefore, be considered a valid approach to automatically detecting scenic routes.

LiDAR point clouds such as AHN2 and AHN3 are increasingly available at city, regional and national scales [13]. The 3D view of the LiDAR data makes them particularly useful for detecting vegetation, as the geometric properties of the LiDAR points and their neighbors can be used to first classify vegetation, and further classify low vegetation and trees based on the height difference.

There are two important steps in classifying LiDAr data: (i) feature extraction and (ii) classification using a machine learning algorithm. Feature extraction is the process of getting additional information of each point. We did that by looking at how the point is located in relation to its surroundings. This is done by defining a neighborhood of a point (for example the closest 15 points) and quantifying the spread of points using a structure tensor. With these extra features we will train a machine learning algorithm. [14]

5. PROJECT SETUP

The project set up describes the processing pipeline and steps taken to meet the research question objectives and find popular scenic routes within the 1.6 million MyDrive routes. The overall pipeline consists of analysing the data through exploratory analysis, using the TomTom Basemap data to coarsely filter out relevant categorized roads that fit within a tourist/scenic road category, analysing the route data for normalised density of planned routes to create a popularity metric, and finally, matching the filtered scenic tourist routes with the aggregated/density clustered routes in order to discover the most popular planned to routes users have planned along scenic roads.

In addition to the above project set up, an experiment was conducted on the LiDAR point cloud data to automatically detect vegetation, as a process to automatically locate scenic routes. A wide variety of tools were utilised for analysis including Spark with both Java and Scala, GeoSpark, Babylon, Tippecanoe and Python (using the GeoPandas, scikit-learn, numpy, Matplotlib and Pandas libraries).

5.1 LiDAR Experiment Set Up

The project is setup by selecting a small study area. 3 shows the study area selected. This is an area of 1.8 square million meters located near Beesd in The Netherlands. The area was selected by searching for an area with an agricultural landscape and vegetation by looking on satellite imagery. A random forest algorithm was selected. Training and testing data was created by manually segmenting areas of vegetation and irrelevant.

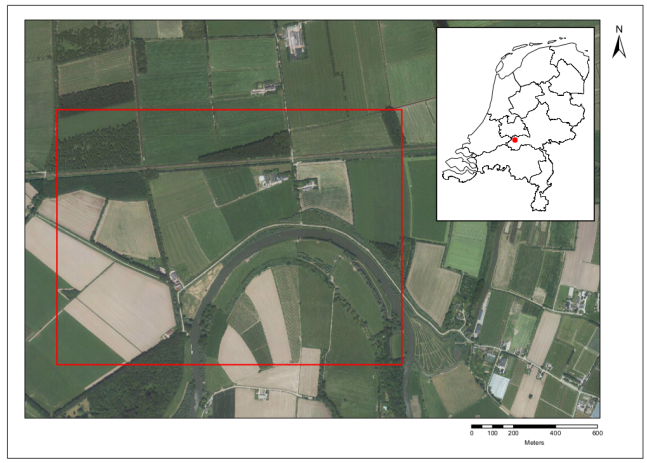


Figure 3: Study area using AHN3 data in The netherlands

The parameters of the random forest (max depth, max number of features, min sample per leaf and min sample per split) was optimised using a cross validation grid search. To assess the accuracy of classification a confusion matrix will be used. This matrix will show the predicted and actual classes of the tested pixel/points in order to give an overview of performance and quantify the errors as well as precision and recall. [16]

5.2 Extracting scenic segments

Due to the fact that we have access to private data from the core maps of TomTom, we decided to create a dataset by extracting stretches of road and routes from the maps based on enriched features that are recorded on the maps. For example. the maps indicate stretches of road that are coastal, or contain mountain passes, forests and waterways.

A scenic route is a stretch of road in a beautiful countryside, often accompanied by panoramic views. Typical examples are coastal roads or mountain passes. The routes within the scenic, regional and cultural category . The other types are genuine routes that were considered relevant as tourist/scenic or aesthetic routes that are more interesting than regular roads . Scenic and nature routes are generally smaller in size and do not have a name.Details on how routes were selected for each category are detailed below.

Scenic Route A stretch of a road that is generally recognized as picturesque due to panoramic views, these are extracted stretches of road that are coastal, mountainous and excludes main roads;

Cultural Route A route connecting certain important cultural or historical sites - e.g. Bier und Burgenstrasse. All marks with historic/cultural/famous features were selected to create this class.

Nature Route A route connecting or running through certain nice natural surroundings, mountains, forests, national parks. These are extracted stretches of road. The difference between scenic and nature is that nature is a stricter category only containing nature elements such as mountains, forests, national parks and water.

Regional Route A real route within a country that is running through a region or connecting region-specific places of interest - e.g. Circuit des Vosges du Nord.;

National Route A route connecting two destinations or connecting national highlights - e.g. Deutsche Ferienroute Alpen Ostsee.

5.2.1 Extracted Segments

Within our definition, an aesthetic or scenic route is a road, or is a chain of road parts that is preferable for a particular reason. This can be:

1. A preferred route from one point to another due to the environmental features.
2. Leading from one interesting place to another,
3. Running through a specific region, or along nice natural surroundings.
4. Routes that carry a name referring to a specific theme or subject, which is the origin of such a route. For example: "Deutsche Fachwerkstrasse".

While this process of extracting aesthetic, scenic and interesting routes based on enriched map features, this process does not give us any granularity on which route may be worth recommending to a user. It only provides a first filtering step.

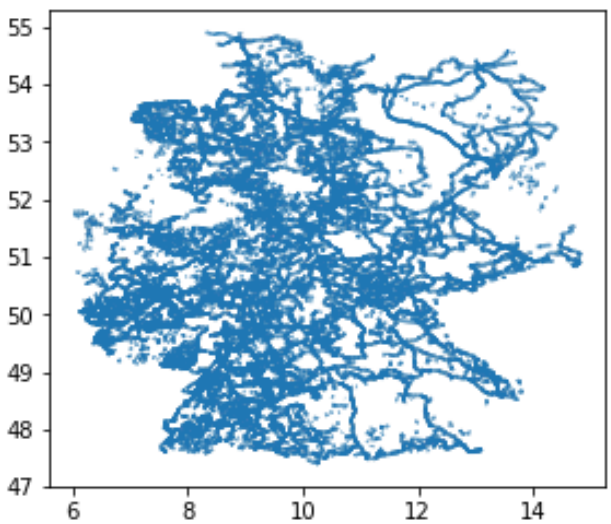


Figure 4: All Scenic Routes in Germany (source *TomTom Base Maps*)

5.3 Route Popularity

5.3.1 Data Pre-Processing

Following exploratory analysis, the following pre-processing and filtering was conducted on the TomTom itinerary routing data.

1. The data format is not correct GeoJson, which required additional filtering in order to parse the coordinates correctly from the Json output. The routes were stored without a correct FeatureCollection for coordinates as per the GeoJson standard.
2. Remove routes that contain 'Home' and 'Work' planned with the 'fastest' cost function for route planning. These routes are unlikely to be scenic or tourist roads.
3. Remove routes planned with the 'truck' option. Trucking routes are limited to driving on major roads that are truck friendly. The aesthetic roads we're looking for are unlikely truck friendly, so the decision was made to exclude these routes
4. Filter routes that contain a hard stop with LocationInfo, which contains POI (Point of Interest) information which is categorized (major tourist attraction, hotel etc.). Unfortunately, only 1000000 of the 1.6 million planned itineraries contain any location info. This filtering was used later for experiments with point of interest ranking.
5. Deal with missing data issues: several routes uploaded from GPX files contain 0, 0 longitude and latitude coordinates. All routes with no, or 0, 0 coordinates were removed from the data.
6. Extracting and filtering out the GPSies.com itineraries as previously described within the Gpsies subsection with section 4
7. Filtering the data by geospatial coordinates within a particular country in order to process each country separately. The reason for doing so is to address the issue that the data is unevenly distributed per country and we decided to normalise per country.

5.3.2 Clustering and Aggregation: Popularity Ranking

Since the MyDrive route itineraries cover almost the entire road network, and the extracted scenic/interesting routes covers a lot of roads as well, it become clear that the popularity and density of planned routes is a valuable resource to utilise in order to identify the popular scenic segments of road. In addition. due to to the fact that the data distribution per country is uneven, we needed to normalise the density clustering analysis per country or region (eg Germany and Scandinavia).

Within the data set, there's two features that are particularly useful to generate a popularity ranking that we could use to find the most popular scenic routes. Specifically, these are *waypoints* and *paths*. A *waypoint* is a coordinate pair which is associated with a stop someone has planned along a route, or the final destination. *Paths* on the other hand are coordinate pairs that make up a GeoJSON linestring that represents a route. A single route has a least one *waypoint*, but may have hundreds of *paths*. In developing a ranking score, we gave preference for the stops users made in the form of a *waypoint* coordinate pair by multiplying the final score result for each *waypoint* coordinate pair by two before merging the final files. In this way, *waypoints*, and the possible points of interest that they represent could also be identified.

5.3.3 Clustering and aggregation

Following the data pre-processing steps outlined in 5.3.1, the data is split into JSON files containing the *waypoints* and *paths* for each country/region.

Within the JSON file, any array of two or more numbers will be treated as a longitude-latitude pair. This includes GeoJSON Points as well as the points that make up GeoJSON MultiPoints, LineStrings and MultiLineStrings. Care was taken that no other numbers would be mistaken for a longitude-latitude pair.

In order to provide a ranking metric for the 1.4 million routes the process was as follows:

1. Using the GeoSpark library we would like to find all the *waypoints* and *paths* within a particular area
2. Leverage the *SpatialRangeQuery()* provided by GeoSpark to return all the *waypoints* and *paths* in a particular region Eg Netherlands
3. Create a spatial KNN (K-nearest neighbour) query, which contains two phases of selection and merging. It takes the resulting partitioned SRDD from (1) and a point P and number K as inputs [18].
4. After the selection phases, we construct k-means clusters in order to reduce the densities and count points. GeoSpark merges results from each partition to the nearest K elements that have the shortest distance to P. the nearest k elements that have the shortest distances to P and outputs the result [17].
5. We then compute the z-order index of each point, the same index that is utilised to divide maps into tiles. The densities can then be manipulated along a linearizion of the point as opposed to computing it within high dimensional space
6. the purpose of computing a z-order score is for visualisation of the popularity densities. Points that are about a pixel apart at zoom level n on screen are around $2^{(64-2*(n+8))}$ apart [20] . The end result is instead of a continuous density function, you have a series of discrete points with an accompanying density score. With this score we perform our match and filter using our scenic ground truth as will be discussed in section 5.3.5

5.3.4 Results of clustering: popularity aggregation

A byproduct of performing the analysis per country is that we end up with popularity of planned routes per country, allowing us to see where users from each country plan routes to. The results of this are interesting, and an example of this is shown below. The routes were separated by specifying a spatial query with waypoint within a particular country. The example below in 5 demonstrating this in the case of The Netherlands.As can be seen in the figure, the data shows that Spain and even Turkey are popular destinations for people planning routes from the Netherlands.

5.3.5 Scenic Matching and Filtering

We propose a matching problem as follows: given a file of GeoJSON objects representing paths and a file of annotated coordinates, we wish to output a file of all points, with annotations, from the second file which appear at least once in any path in the first file. To this end, we first transform the GeoJSON file into an RDD of points. Since Spark has native support for files where each line is a JSON object we can immediately load the GeoJSON objects into a dataframe.

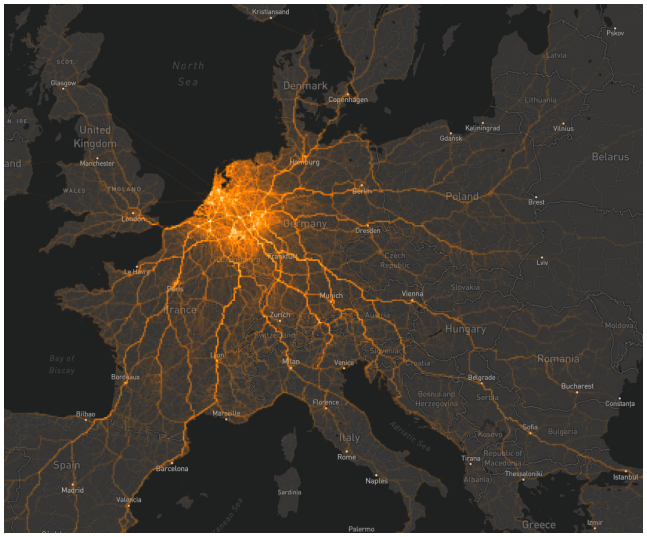


Figure 5: Popular routes in The Netherlands

This ensures that a column exists which contains an array-like type containing the points. Using the *explode* command we can then expand these arrays to create a new dataframe where each row contains a single point. By selecting a subset of columns from this dataframe and applying a set schema we can achieve a program state where we have access to two datasets containing one spatial point per line.

To match points between these two datasets, it may be tempting to perform a simple SQL inner join when both the latitude and longitude of two points are equal. However, due to the precision of the spatial coordinates of the dataset, it is highly unlikely that two measurements of approximately the same point in space will have exactly the same latitude and longitude. Thus, we must perform a fuzzy inner join on the two datasets. We propose three different ways of performing a fuzzy join on coordinate data; we can either (a) use a library such as GeoSpark to determine approximate equality, or (b) join on multiple conditions using calculated minimum and maximum latitudes and longitudes, or (c) round the latitudes and longitudes with some level of precision and perform an exact join on the rounded value. We believe that the latter option provides a good compromise between performance and simplicity; we round coordinates such that they uniquely represent points at distances of approximately 10 meters.

The end result shown above in figure 6 is visualised to show the density of popular routes in the same manner discussed previously in 5.3.3. The higher density points have a brighter colour to represent them and show the most popular locations of planned routes within the scenic nature areas, in this case. The same was process was completed for each category as previously defined in **??**

5.4 Visualisation

As a final visualisation result we produced an interactive visualisation of the popularity of routes across the defined aesthetic categories, namely nature, scenic, cultural, national and regional as detailed in 5.2. The visualisation contains a base map of 1.4 million European itineraries with density clusters in order to visualise the density of

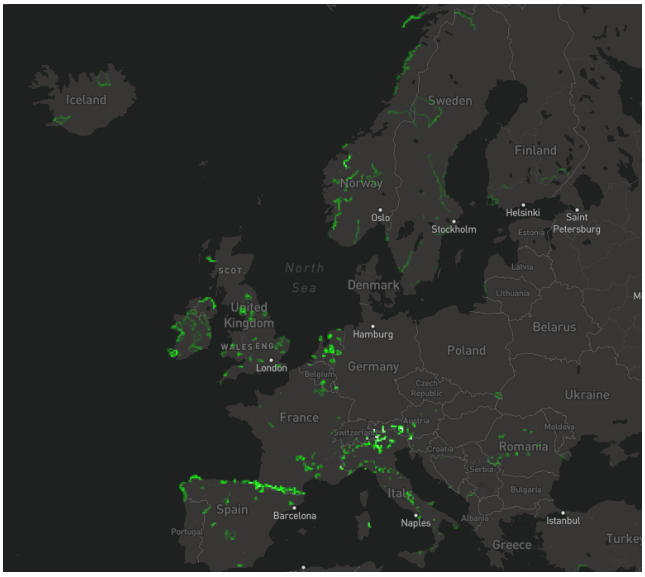


Figure 6: Scored Nature Route Matches in Europe

routes. We then allow comparison of the 1.4 million European itineraries with the matched aesthetic routes which can be selected by the user. The user selects an aesthetic category and then is able to drag across an object to perform an X-Ray like comparison between the clusters of all European routes, and the clusters of popular routes within a scenic category that was the result of our analysis. We also included a GPS coordinate when hovering over with the mouse to allow easy lookup of any area on the map.

We utilised a Javascript library called MapBox GL JS to produce the maps, with data stored in an MBTiles format, which is a file format, or technically a SQLite database for storing map tiles. The MBTiles are stored in MapBox Studio to provide a lightweight responsive static HTML page.

6. EXPERIMENTS RESULTS

In this section we briefly explain the experiments results. We ran all our experiments on a TomTom OpenCloud Hadoop Cluster and on the SurfSara Hadoop cluster of 90 machines, 720 cores and 1.2PB of storage as well as our personal laptops.

6.1 LiDAR Classification Experiment Results

After performing classification of vegetation, the results of the test is shown in figure 7. The accuracy of the classifications are presented in a confusion matrix (table 1). We also show the measure for unbalanced data sets in table 2. [15]

The ROC-AUC of 0.95 demonstrates that the two classes classified separately well. This is further indicated by the Matthews correlation coefficient of 0.59, indicating a positive correlation between predicted and observed classes. The confusion matrix shows a recall of 0.97 for the vegetation class and 0.72 for the non-vegetation class.

Based on 8 which shows the different features, the most important feature for the model is the number of returns feature when filtering out vegetation. This is due to the fact that vegetation is the most common object scanned which produces multiple returns from the ariel LidAR scanning

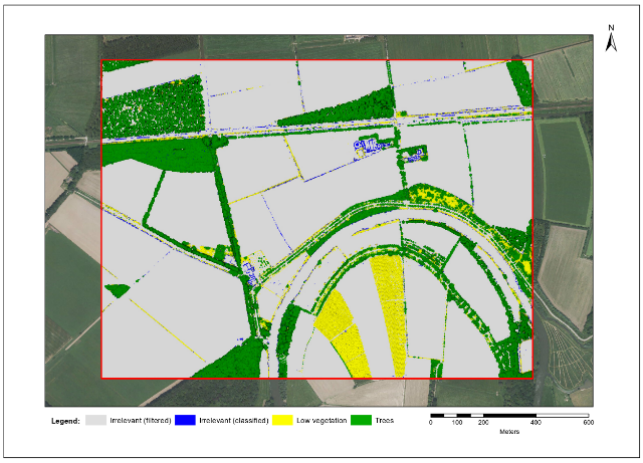


Figure 7: A map of the classification results.

Table 1: A confusion matrix showing the predicted classes against the actual classes of the 10 fold cross-validation accumulated.

Predicted	Vegetation	Irrelevant	Recall
Actual			
Vegetation	1159762	23438	0.97
Irrelevant	6416	16706	0.71

Table 2: Assessment of the accuracy using the average ROC-AUC, MCC and geometric mean of the 10 fold cross-validation.

Metric	Score
ROC-AUC	0.95
MCC	0.61
Geometric Mean	0.83

process. It is, however, not the only objects which produce multiple returns. The power line for the railway produces a high multiple returns rate. Conversely, on the edges of vegetation boundaries there's only a single return. The other features are used in combination to train the random forest algorithm.

While this method and the AHN3 data is suitable to classify vegetation on samples of data, it was not successful enough to be a viable method to classify scenic routes due to the limited data coverage of AHN3, and the fact that there is very limited open, high quality liDAR data across Europe.

The method could, however, be used as a feature that, along with several additional features, could train a machine learning model to classify scenic routes. The AHN3 data would be better as it includes a number of features that are not included in AHN2 data, however, the AHN3 data is incomplete. The most important two being multiple returns and intensity information. Particularly multiple returns are very useful when classifying vegetation as shown above as vegetation often causes multiple returns when scanning with a laser. [15]

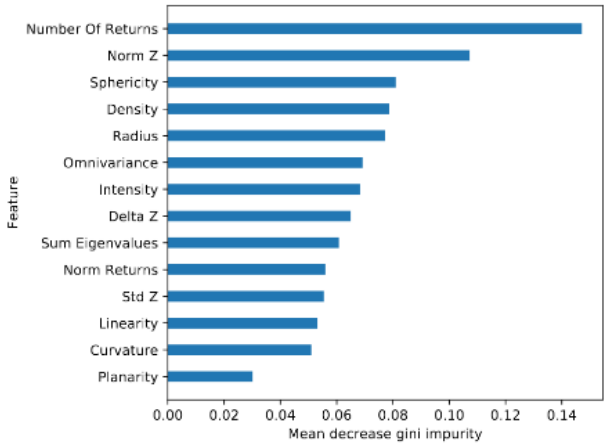


Figure 8: The importance of each feature in the classifier, based on the mean decrease in impurity by the feature. Higher is more important.

6.2 POI Ranking Experiment

As an additional experiment, we tested to see if it would be possible to rank points of interest or POIs by using our count metric derived from waypoints. For example, in order to locate the most popular hotels or restaurants. Using the category search API a search was made within the Noordwijk region and 100 hotels were retrieved from the 'hotel' category and the same match and filter process was applied against the aggregated route data. This is shown below in 9.

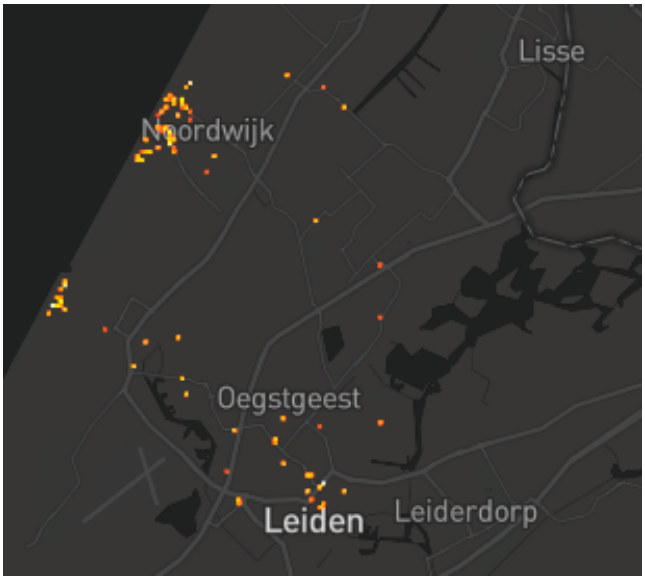


Figure 9: 100 Hotels ranked in Noordwijk

While this approach provides a ranking for hotels it is not a thorough approach and is limited by several factors. Firstly, the data was not separated by seasonal periods which can be a large factor in a particular hotels popularity. Secondly, popularity of planning a route to a hotel may be affected by the presence of other points of interest on the same property, such as restaurants. Additionally, there's no information about the quality of the hotel, or star rating. We believe this score ranking approach could be used in combination with other data sources to train a supervised learning to rank algorithm. [19]

6.3 GPSies.com Aggregation

As previously discussed in section 4.3, a subset of routes was found within the data during data exploration and we aimed to aggregate this data separately in order to re-rank the data, as we lost all granularity of the gpsies routes when the GPX files were uploaded and saved into our dataset.

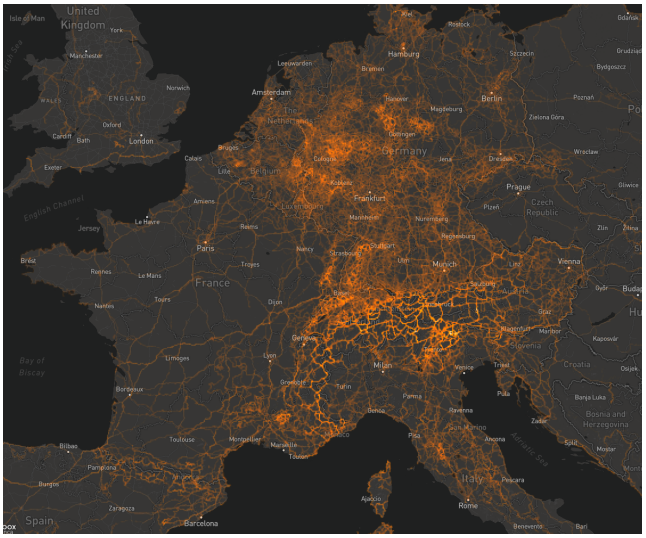


Figure 10: Heatmap of GPSies.com routes with score

The heatmap plot shown in figure 10 shows the results of the aggregation of the gpsies.com data. We were able to create density clusters based on the saved routes. These routes themselves could be a good source of scenic routes due to the fact that the website is dedicated to finding worthwhile routes to explore and has a star rating for each route. We did not investigate our results further as this was a side experiment and out of the scope of our main project set up for addressing the research questions.

6.4 Evaluating Scenic Route Classification Results

Our main project set up was completed in order to attempt to classify scenic routes across different categories, and find the most popular scenic routes in order to address our research questions. Our conclusion and assessment of meeting our research questions is discussed below in 7. Within this section we evaluate the results of our scenic classification.

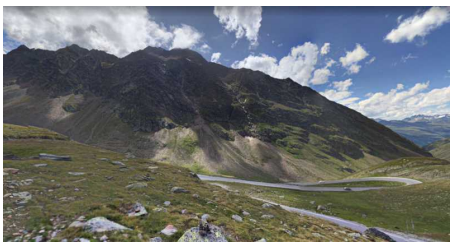
The top ten scored results within each of the categories, namely: scenic, nature, cultural, national and regional is included within the appendix 7.1. The latitude, longitude and score is included for each category. These represent the pair of coordinates with the highest score after performing the filter and matching algorithm to match the overall route popularity and the extracted scenic segments. A 2D histogram heatmap of the results id shown in 13

We used Google Maps and TomTom streetview services in order to visualise and validate these coordinate pairs with real life photos of the points. We display the top three points from the *nature* category as shown below 11 as well as the top three results in the *culture* category 12 in order to show our positive and negative results. The *scenic* category is similar to our results from *nature* , and *national* similar to the results from *cultural*.

The results demonstrate that the process worked in matching popular and scenic segments, however, the results in terms of their scenicness are subjective. In our opinion the results from the nature category are compelling, and it's clear why they're popular areas for routes being planned. The culture category results are less appealing, primarily due to the fact that while the roads are culturally significant in some way, they are also popular main roads used to travel on.

(a) Via Strada Statale 38 dello Stelvio, Lombardia, Italy (source: *TomTom Cartopia*)

(b) Ötztalstraße B186, Tirol, Austria (source: *TomTom Cartopia*)

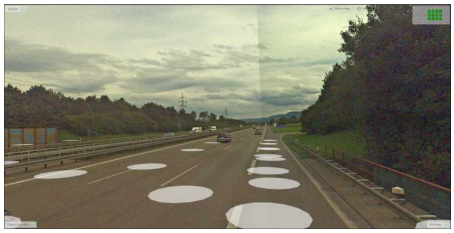


(c) Elsewhere on the Ötztalstraße B186, Tirol, Austria (source: *Google Streetview*)

Figure 11: The three most scenic routes in the nature category according to our algorithm.

7. CONCLUSIONS

Using the data and approach we were not able to fully answer our research questions, although we did partially address them. We were not able to find which routes are scenic within our itinerary routes and automatically classify them, but we were able to find scenic routes within our data as well as create density clusters of the most popular routes and apply these to scenic routes. Our experiments with



(a) Bundesstraße 297 near Kirchheim unter Teck, Baden-Württemberg, Germany (source: *TomTom Cartopia*)

(b) Bundesautobahn 61 near Waldorf, Rhineland-Palatinate, Germany (source: *TomTom Cartopia*)

(c) Bundesautobahn 61 near Mendig, Rhineland-Palatinate, Germany (source: *Google Streetview*)

Figure 12: The three most scenic routes in the cultural category according to our algorithm.

POI ranking and using LiDAR point cloud data show the data could yield a result given additional data and extensive model training. In our view, the result we achieved given the research time constraints provides moderate success at addressing the research questions. We have successfully created popularity clusters for over 1.4 million itineraries within Europe and applied this to a set of extracted scenic routes.

As the results show in section 6.4, the top rated scenic and nature locations are within what most would consider very scenic and aesthetic locations. Although we admit this is a subject question, our top results feature mountainous panoramic views with trees and nature primarily located in the popular greater Alps region. Specific routes are also found in the data, such as the Keukenhof flower route in The Netherlands which shows up as one of the most popular routes within the Regional category.

We also note a few problems with our results. The solution is only scalable given enriched map data to extract ground truth, as well as planned itineraries within the region you're assessing. The same is true, however, for other approaches. Using LiDAR, satellite or images to classify scenic routes would require high quality data sources in all regions and adequate training data which would require manually labeled images. We also note that training a model for identifying scenicness will not generalise well. Scenic routes in Spain do not look the same as in Russia for example.

Additional problems with our results is that the cultural

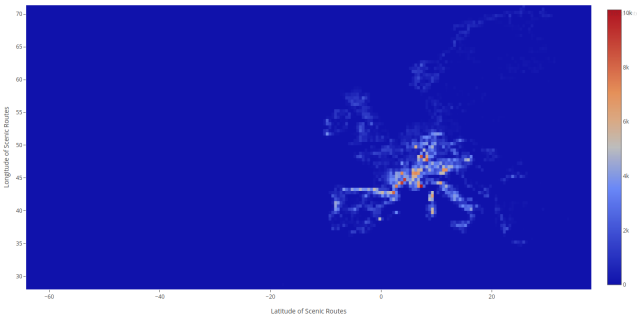


Figure 13: 2D Histogram Heat Map of Scenic Routes

and national category are arguably not scenic at all. While they represent some important route such as an old historic Roman road, or Route 66 in the US, this ends up being very different to a panoramic ocean road. This is one of the reasons we separated the analysis and results into different categories and we believe the cultural category still represents interesting routes compared to for example, a typical "fastest" A* algorithm generated route.

Lastly, there's problems with the density results around scenic routes that may also be main roads, as main roads may be more popular due to factors that are not aesthetic.

7.1 Future Work

We have several suggestions of improvements for future work. Firstly, within the context of the methods we have utilised, normalisation of main roads intersections where their is high density would be a first step. Second, separating the results into seasons could provide a more granular result showing popular routes in the summer, spring and winter independently. The same approach could be used for extending the POI ranking experiment that was conducted.

While there's limitations with using images, or LiDAR as was discussed above within section 7, our suggestion for future work would be combining these different approaches, deriving features from each approach and training an ensemble based model that assesses scenicness based on the set of derived features. An example could be, map features, LiDAR and satellite images feature, user photos and popularity. Based on a combination of these derived features an ensemble tree-based machine learning approach could be used to automatically classify and rank scenic routes.

Due to the lack of time, our approach is limited and the features we use to classify scenic roads lack automated scalability and heavily uses enriched map data. Despite this, we do provide some interesting results as well as unexpected discoveries within the data such as the GPSies data and finding popular routes that different countries plan routes to, and using the popularity for point of interest ranking. We also found some genuinely scenic routes from our top results within the scenic and nature category.

References

[1] Majid Alivand, Hartwig Hochmair, and Sivaramakrishnan Srinivasan. "Analyzing how travelers choose scenic routes using route choice models". In: *Computers, Environment and Urban Systems* 50 (Mar. 2015), pp. 41–52. DOI: 10.1016/j.compenvurbsys.2014.10.004.

[2] Yu Zheng. "Trajectory Data Mining". In: *ACM Transactions on Intelligent Systems and Technology* 6.3 (May 2015), pp. 1–41. DOI: 10.1145/2743025.

[3] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. "The shortest path to happiness". In: *Proceedings of the 25th ACM conference on Hypertext and social media - HT'14*. ACM Press, 2014. DOI: 10.1145/2631775.2631799.

[4] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. "A demonstration of GeoSpark: A cluster computing framework for processing big spatial data". In: *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, May 2016. DOI: 10.1109/icde.2016.7498357.

[5] J. Boehm, K. Liu, and C. Alis. "Sideloading – Ingestion of Large Point Clounds Into the Apache Spark Big Data Engine". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B2 (June 2016), pp. 343–348. DOI: 10.5194/isprs-archives-xli-b2-343-2016.

[6] Kun Liu and Jan Boehm. "Classification of big point cloud data using cloud computing". In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 40.3 (2015), p. 553.

[7] Lars E. Olsson et al. "Happiness and Satisfaction with Work Commute". In: *Social Indicators Research* 111.1 (Feb. 2012), pp. 255–263. DOI: 10.1007/s11205-012-0003-2.

[8] Raymond W. Novaco and Oscar I. Gonzalez. "Commuting and well-being". In: *Technology and Psychological Well-being*. Ed. by Yair Amichai-Hamburger. Cambridge University Press, pp. 174–205. DOI: 10.1017/cbo9780511635373.008.

[9] Jeffrey C. Hallo and Robert E. Manning. "Transportation and recreation: a case study of visitors driving for pleasure at Acadia National Park". In: *Journal of Transport Geography* 17.6 (Nov. 2009), pp. 491–499. DOI: 10.1016/j.jtrangeo.2008.10.001.

[10] Nina Runge et al. "No more Autobahn!" In: *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI'16*. ACM Press, 2016. DOI: 10.1145/2856767.2856804.

[11] Matei Zaharia et al. "Spark: Cluster computing with working sets." In: *HotCloud* 10.10-10 (2010), p. 95.

[12] J Böhm et al. "The Iqmulus urban showcase: Automatic tree classification and identification in huge mobile mapping point clouds". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*. Vol. 41. International Society of Photogrammetry and Remote Sensing (ISPRS). 2016, pp. 301–307.

[13] AHN. *Inwinjaren AHN2 & AHN3*. [Online: accessed October 2017]. 2016.

[14] AS Antonarakis, Keith S Richards, and James Brasington. "Object-based land cover classification using airborne LiDAR". In: *Remote Sensing of Environment* 112.6 (2008), pp. 2988–2998.

[15] Pierre Baldi et al. "Assessing the accuracy of prediction algorithms for classification: an overview". In: *Bioinformatics* 16.5 (2000), pp. 412–424.

[16] Li Guo et al. "Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.1 (2011), pp. 56–66.

[17] K Nafees Ahmed and T Abdul Razak. "A Comparative Study of Different Density based Spatial Clustering Algorithms". In: ().

[18] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. "Geospark: A cluster computing framework for processing large-scale spatial data". In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2015, p. 70.

[19] Xiaoyan Cai and Wenjie Li. "Ranking through clustering: An integrated approach to multi-document summarization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.7 (2013), pp. 1424–1433.

[20] Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. "Mining travel patterns from geotagged photos". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), p. 56.

APPENDIX

Top Ten Nature Scored

latitude	longitude	score
46.507727	10.366802	17948
46.904542	11.091042	16741
46.910406	11.089153	16448
46.504537	10.360107	15365
46.895276	11.046753	13325
46.509971	10.375385	13019
46.537964	10.433750	12799
46.935613	11.033535	12661
46.530997	10.460873	12316
45.870888	10.875950	12282

Top Ten Scenic Scored

latitude	longitude	score
47.352199	10.841618	27437
47.352315	10.841446	20996
46.833187	11.171036	18960
46.330099	11.274548	18922
47.343708	10.818615	18826
46.833187	11.166744	18169
46.507727	10.366802	17948
48.641190	9.430561	17803
51.378317	12.184868	17578
47.322069	10.824623	17489

Top Ten Cultural Scored

latitude	longitude	score
48.641190	9.430561	17803
50.473348	7.227116	15456
50.390791	7.252693	15433
50.593045	7.030048	13754
47.396955	0.707760	12965
46.628928	10.762653	11574
49.257386	3.961601	11551
49.517073	11.305103	11032
49.477159	11.176529	10343
46.627867	10.769520	10323

Top Ten Regional Scored

latitude	longitude	score
48.643798	8.028774	15080
48.957003	8.383770	14517
45.873756	8.976345	13749
47.803931	12.297134	13501
50.314229	7.493019	12864
52.775563	9.667969	12810
53.005178	8.701000	11504
47.517433	10.408344	10961
47.815345	12.363224	10818
53.038736	8.898926	10609

Top Ten National Scored

latitude	longitude	score
46.330099	11.274548	18922
51.378317	12.184868	17578
46.120726	11.085892	14730
48.957003	8.383770	14517
45.900880	11.025639	14033
43.922503	7.407532	13976
44.889931	10.085106	13541
44.802058	6.734104	13211
46.197775	11.128979	12972
45.164248	6.424427	12555