The ORVal Recreation Demand Model: Extension Project

Authors: Prof Brett Day, University of Exeter
Dr Greg Smith, University of Exeter

Date: 15th September 2017

Contents

List of Tables

List of Figures

1. Introduction

The original Outdoor Recreation Valuation (ORVal) project set out to build a recreational demand model that could be used to estimate the recreational welfare value derived from any existing or new greenspace in the whole of England. While the methods of recreational demand modelling are well established, the previous extent of application has tended to focus on just one form of greenspace (e.g. beaches, lakes or municipal parks) and been limited to a restricted geographic region. The challenge of the first ORVal project was to show that it was possible to generate a recreational demand model that functioned at the scale of an entire nation and encompassed the full spectrum of outdoor greenspace localities. Moreover, the project sought to develop an online tool that could be used to interrogate this recreation demand model and explore the values that are or could be generated by existing or new recreational sites.

Phase II of the development of the Outdoor Recreation Valuation (ORVal) Tool, has sought to move from the prototypes of the original analyses to a full working model and tool. A key part of that undertaking has been a major programme of empirical work resulting in a significantly extended and improved ORVal Recreation Demand model. At the outset, the objectives of the empirical analysis were as follows;

- To supplement the estimation data on the outdoor recreational choices of English residents with data from the Welsh Outdoor Recreation Survey,
- To extend the model to include the choice as to which mode of transport to use when undertaking a recreation trip to outdoor greenspace.
- To calculate precise travel costs for driving and walking to greenspaces based on detailed road and path networks for the UK augmented by precise calculations of fuel consumption and the latest research on the cost of travel time.
- To extend the characterisation of decisions over participating in outdoor recreation; particularly in allowing for differences in participation across ethnic groups and to explore the impact of differing weather conditions on recreation participation.
- To extend the characterisation of choice of recreation site to allow for various details of the quality of greenspace including, river water quality, beach type, bathing water quality and woodland type
- To carry out a programme of specification exploration and model testing in order to identify the most appropriate model to use in the tool and document its merits relative to other specifications.

The purpose of this report is to document this programme of empirical work.

2. Recreation Demand Modelling

The approach economists normally adopt to estimate the welfare derived from a good is to observe how demand for that good changes as its price changes. In essence, that relationship traces out how much money individuals are willing to give up in order to enjoy that good; a quantity that (roughly speaking) defines the measure of welfare that economists call economic value. Indeed, throughout this report when we talk about 'value' or 'valuation' we are referring to this particular monetary measure of welfare.

More often than not, however, access to greenspaces does not command a price, or if it does that price is often minimal and without sufficient variation to directly estimate the demand-price relationship. Hence, conventional techniques of welfare estimation are frequently not applicable to the valuation of greenspace. A solution to this problem was first forwarded by Harold Hotelling in a letter to the National Park Service of the United States in 1947 (Smith and Kaoru 1990). He noted that though the greenspace is not itself a market good, in undertaking a recreational trip individuals incur time and travel costs that in effect can be considered the 'price' of access. In other words, when we observe an individual taking a trip to a greenspace, we can presume that the value they derive from that experience is worth at least the costs incurred in travelling to the site.

When considering just one site, this travel cost method progresses by examining how many trips individuals living at different distances, and hence with different travel costs, choose to make to the recreational greenspace. Information of that nature is sufficient to inform on the value for that particular site. The challenge for the ORVal project was considerably different from the single site case. In particular, we were concerned with recreational activities over all greenspaces in England where those greenspaces were differentiated not only in their location but in the recreational experience they offered.

A related framework that better suits our needs is one that focuses on an individual's choice of which of the array of different greenspaces to visit rather than how many trips to take to a particular greenspace. This discrete choice approach is also a form of travel cost modelling. The intuition of how information on discrete choices provides evidence for welfare valuation progresses as follows. Imagine, an individual has a choice between just two greenspaces. Both greenspaces provide visitors with 2ha of open grassland but the more distant greenspace also possesses 2ha of woodland. If we observe the individual choosing to visit the more distant greenspace we can conclude that the extra welfare derived from being able to visit a greenspace with woodland must be worth at least as much as the extra costs in travelling to that more distant location rather than the closer greenspace. Given sufficient observations on individuals choosing between quality-differentiated greenspaces at different distances from their homes, the discrete choice approach can inform on the economic value that individuals realise from greenspaces with different qualities. Moreover it can be used to predict how likely it is that an individual will choose to visit a particular greenspace from the set of greenspaces available to them.

The econometric method used to estimate discrete choice models are known as Random Utility Models (RUMs). We review the particular RUM approach used in the ORVal empirical analysis model in Section 4. The approach is data intensive. It requires information on the choices individuals make on each recreational choice occasion (in our modelling we assume that each day represents such a choice occasion). In particular, we need to know whether an individual took a trip to greenspace or not and, what mode of transport they decided to use in getting to that location, what the qualities of that site were and the time and travel costs incurred in getting there. Moreover, since this is a choice model, we need details of the qualities associated with each other recreational greenspace that individual might have visited instead and the travel costs of different modes of transport associated with reaching each of those alternative locations. In Section 2 we describe the data sources used to construct such a data set then in Section 3 how that data was processed to generate the estimation data set. Finally, Section 5 describes the modelling results and model specification testing.

3. Data

3.1 MENE Data

The primary data set supporting estimation of the ORVal model is provided by the Monitor of Engagement with the Natural Environment (MENE) survey. Administered on behalf of Natural England, DEFRA and the Forestry Commission, the MENE survey provides a large, random location sample of recreational day trips taken by adults (over 16 years of age) residents of England. As a consequence, the estimates of visits and values that are estimated from the ORVal model are limited to:

- Recreational day trips
- Residents of England
- Adults (over 16s)

The survey is administered face-to-face, recording the recreational trips to greenspace taken by the respondent over the seven days prior to the interview. Moreover, for one randomly selected trip, the survey elicits detailed information regarding the respondent's activities on that trip as well as the location and characteristics of the recreational site visited. In this report we describe this trip as the *focus visit*.

The MENE survey runs throughout the year sampling at least 800 respondents each week making the data seasonally representative. As recorded in Table 1, the annual sample amounts to approximately 50,000 respondents. The ORVal extension project took advantage of the release of the 2015-16 survey release meaning that the modelling was based on seven years of data collected since the survey began in 2009.

Table 1: Annual sample sizes in the MENE survey

Year	Sample
2009-10	48,514
2010-11	46,099
2011-12	47,418
2012-13	46,749
2013-14	46,785
2014-15	45,225
2015-16	45,965
Total:	326,755

The MENE data is provided with a demographic weight for each observation. The weight is calculated so as to ensure that the sample of respondents collected in one month can be adjusted so as to be representative of the adult population of the UK in that year. The demographic characteristics used in calculating the weights are:

- age and sex (for example, males 16- 24, females 85+),
- region of residence,
- social grade,
- presence of children in the household,
- sex and working status (for example, male full time),
- presence of a dog in the household and
- urban/rural residence

Put simply the weight for each observation indicates the number of people in the population represented by that respondent. Accordingly, the weighted sum of observations of, for example, male respondents aged 16-24 will equal the number of males in that age group in England, with the same being true of all the other demographic categories in the list above.

One of the objectives of the ORVal extension project was to model the choice of whether to walk or travel by motorised vehicle to a recreational site. In the original modelling effort we had assumed that all travel was by car. Under that assumption we took it as sufficient to model people's home location as being the population weighted centroid of their home LSOA (lower super output area), reasoning that the error in the calculations of travel times and distances from home to a site would be of a scale that was decision irrelevant when travelling by car. Such reasoning would not hold for walking decisions, where one might expect decisions to swing on significantly smaller differences in distance. As part of the extension project, therefore, we were grateful that Natural England (through the survey company Kantar TNS) furnished us with postcode data for respondents in the MENE survey, allowing us to identify a highly accurate estimate of home location.

The MENE survey is a pseud-diary study asking respondents to recall their outdoor recreational activity over the seven days previous to the interview. For one randomly selected visit made over that prior week the survey goes into detail, recording information on the exact location of the greenspace visited and various details of the sites characteristics and the respondents activities. The model described subsequently makes use of this entire diary of recreational activity.

A detailed description of the MENE survey, its administration and the calculation of demographic weights can be found in the MENE Technical Report (Natural England 2015).

3.2 WORS data

One of the original objectives of the ORVal extension project had been to include data from the Welsh Outdoor Recreation Survey (WORS) in the estimation of the recreation demand model. That model is reasonably demanding of data, at a minimum requiring information on the respondent's home location, the location of the site they visited, details of when the trip was made and how the respondent travelled. A key first step in the extension project was to source the WORS data and examine the extent to which the data was compatible with a programme of joint estimation with the MENE data.

Three iterations of the WORS survey have now been undertaken with data available from 2008, 2011 and 2014. While the 2008 and 2011 surveys did not record the locations of recreational visits, extra details were taken in the 2014 survey allowing visit locations to be geolocated in the same way as is done for the MENE survey. Accordingly, only the 2014 lent itself to possible use in model estimation. Again we were grateful that Natural Resources Wales provided us with the site geolocation data as well as home postcodes for the sample respondents.

The WORS survey differs from the MENE survey in one other important manner. While the MENE survey pursues a strict diary response model, collecting data on outdoor trips taken over the seven days previous to the survey, the WORS survey asks about recreation activity over the last month recording information on the last trip taken. For our purposes that posed a problem, in so much as no details are taken as to the date of that focus trip. The ORVaL model uses data on the weather, day of week and time of year as key explanatory variables in the choice of participation. Without knowing the date of the focus trip pulling such data together for the Welsh sample would be very difficult. We hoped that we could get part way to understanding when the trip was taken by using data on the data of interview to at least establish the time of year of the visit. Acting on our behalf NRW approached the survey company that administered the survey to release that information for our use. Unfortunately, it transpired that the survey company had stripped out the survey date records from their stored data such that that information was no longer available.

Ultimately, it was decided that the differences in the MENE and WORS data sets mitigated against a programme of joint estimation. Rather, the revised plan was to estimate the recreation demand model from the detailed MENE data then use a process of calibration to adjust specific parameters of the model to fit with higher level data from the WORS survey. In particular, to introduce a Wales-specific variable to the model and then adjust this parameter until the model's estimates of rates of participation in outdoor recreation amongst Welsh residents best match the rates of participation recorded in the WORS survey in 2008, 2011 and 2014.

3.3 ORVal Greenspace Map

As well as information on recreational trips, a second key dataset needed for estimation of a recreation demand model are details of the locations of sites for outdoor recreational activity. In this project that data were provided by the ORVal greenspace map. The ORVal greenspace map is a detailed spatial dataset compiled through the combination and manipulation of a large number of primary data sources that describes the location and characteristics of accessible greenspace across England. Construction of the ORVal greenspace map is provided in Day (2016).

As part of the extension project, the ORVal greenspace map has been extended to include Wales. An expert of that dataset is shown in Figure 1. The Welsh greenspace map was constructed using an algorithm that replicates the process used to generate the original English data. Briefly, a data set of 'parks' was generated through the collation of the following datasets;

- *Country Parks*: Data sourced from the LLE geoportal.
- *National Nature Reserves*: Data sourced from the LLE geoportal.
- *Local Nature Reserves*: Data sourced from the LLE geoportal.
- *Openstreetmap (OSM) Parks*: A download of features ('ways' and 'relations', in OSM terminology) from the OSM where the tags 'landuse' or 'leisure' are given the values 'park', 'common', 'recreation_ground' or 'village_green'. So as to focus on open access recreation areas, OSM Parks was then tidied to remove features where access is tagged as 'private' or 'restricted', or as being the grounds of a leisure centre or sports club, school playing fields, hospital grounds or supermarket premises. To remove small areas of amenity grassland the

data was cleaned to exclude roadside verges and roundabouts before finally deleting all unnamed OSM parks less than 0.4ha is extent and not containing a playground.

- *OSM Nature Reserves*: A download of features from the OSM where the tags ‘landuse’ or ‘leisure’ are given the value ‘nature reserve’.
- *OSM Public Gardens*: A download of features from the OSM where the tags ‘landuse’, ‘leisure’ or ‘amenity’ are given the value ‘garden’ and where ‘access’ was specifically labelled as 'public', 'yes', 'permissive' or 'destination'.
- *Forestry Commission Recreation Areas*: A data set constructed from the ‘National Forest Estate England Recreation Routes’ dataset.
- *Woods for People*: Data sourced from the Woodland Trust’s ‘Woods for People’ dataset restricted to areas in excess of 0.4ha extent.
- *OSM Cemeteries*: A download of features from the OSM tagged as cemeteries or graveyards. Small cemeteries of under 0.2ha extent were removed from the data.
- *OSM Allotments*: A download of features from the OSM tagged as allotment.

A Welsh ‘paths’ data set was constructed from OSM data by selecting features in which the key Highway had been tagged as ‘track’,‘footway’,‘path’,‘cycleway’, ‘byway’, ‘trail’,‘bridleway’. The data were reduced by removing all features for which access was private or otherwise restricted. Path stretches in urban areas that did not border waterways or other green features were eliminated and an algorithm used to gather paths together into networks of connected paths. Access points to paths were identified by points of intersection between path networks and the roads network.

Finally, a beach location dataset was constructed from the http://britishbeaches.info website. The data were cleaned and merged according to the rules described in the original ORVal data report (Day, 2016). Note that the Welsh data does not contain information on greens created under the Millennium Greens and Doorstep Greens schemes which were restricted to England.

As described in Table 2, the ORVal greenspace map identifies some 138,617 greenspace sites in England and Wales that could form the focus of a recreational trip.

Table 2: Recreation sites in the ORVal greenspace map

Type	Number of Sites	
	England	Wales
Parks:		
Municipal Park	19,363	1,000
Cemetery	8,230	783
Woods	7,359	1,241
Allotment	6,865	198
Nature	2,844	211
Country Park	413	37
Path Access Points	82,591	6,621
Beaches	630	231
Total	128,295	10,322

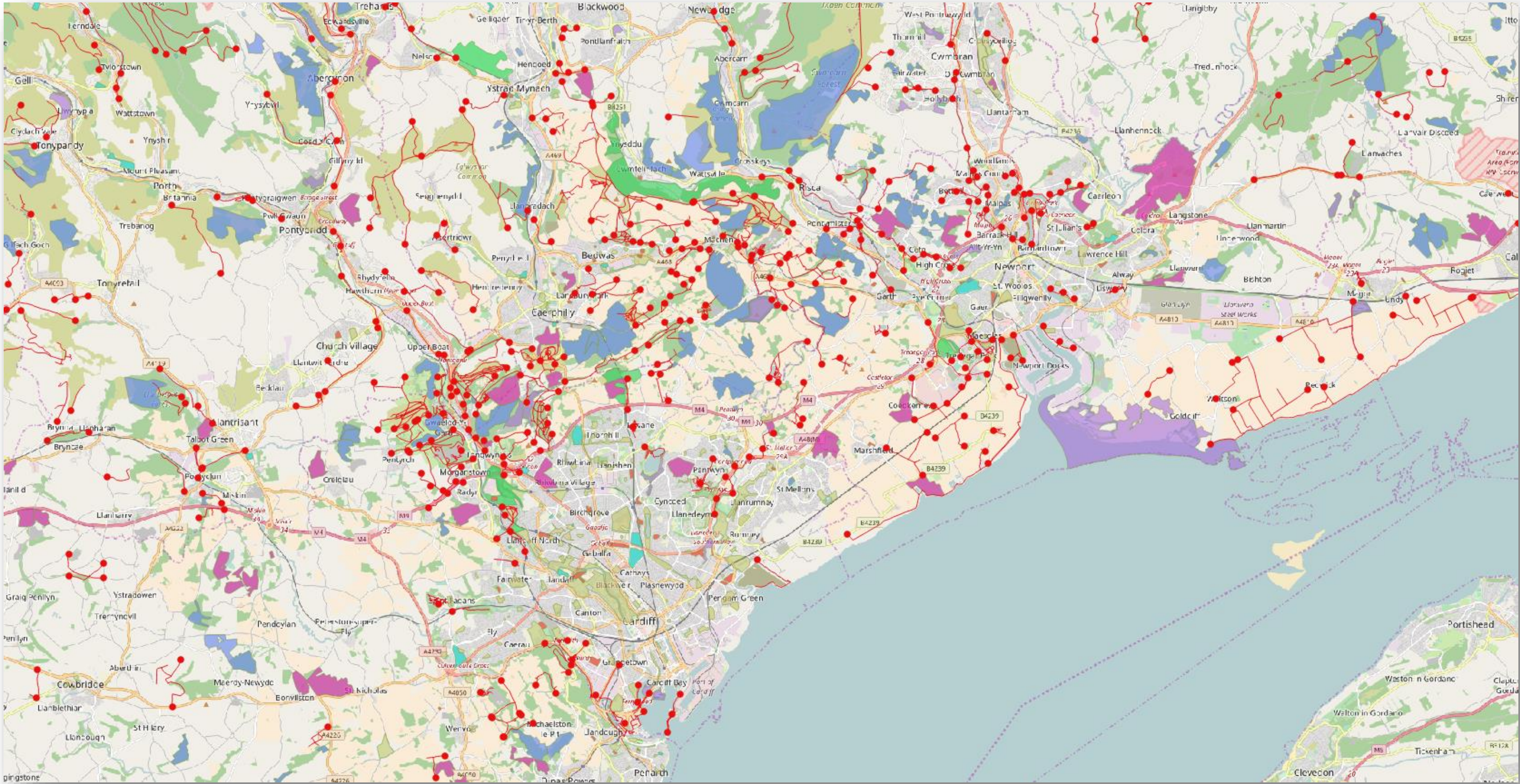

Figure 1: ORVal Greenspace Map for Wales - Cardiff & Newport

The recreation features identified on the ORVal greenspace map come in three basic forms;

- parks which consist of areas of accessible greenspace within well-defined boundaries over which visitors usually have freedom to wander at will,
- paths which consist of accessible, walkable routes that pass through the landscape, often traversing a variety of different greenspaces and tending to restrict visitors to defined routes of passage.
- beaches.

Each recreation site is described by various aspects of its physical characteristics; particularly the site's dimensions, landcovers, designations and points of interest.

Table 3 provides an indication of landcovers used to describe sites and how frequently those landcovers were present at the various sites. Note that sites are characterised by a diversity of land covers so the columns of Table 3 do not sum to the number of sites of different types shown in Table 2. Moreover, for paths the presence of a landcover is determined by whether that landcover was found along the path network accessed by a path access point with in 10km of that access point. Further details can be found in the ORVal Greenspace Map report (Day, 2016).

Table 3: Landcovers present at recreation sites

Landcover	Parks		Paths		Beaches	
	Number	Percent	Number	Percent	Number	Percent
Woods	18,151	37.4%	75,690	84.8%	0	0.0%
Wood Pasture	1,047	2.2%	11,382	12.8%	0	0.0%
Agriculture	439	0.9%	78,301	87.8%	0	0.0%
Natural Grass	4,070	8.4%	64,824	72.7%	0	0.0%
Moors	867	1.8%	17,323	19.4%	0	0.0%
Mountain	26	0.1%	946	1.1%	0	0.0%
Coastal	1,239	2.6%	3,713	4.2%	861	100.0%
Saltmarsh	218	0.4%	2,308	2.6%	0	0.0%
Marsh & Fen	632	1.3%	8,844	9.9%	0	0.0%
Managed Grass	19,245	39.6%	83,343	93.4%	0	0.0%
Sports Pitches	4,177	8.6%	4,600	5.2%	0	0.0%
Gardens	571	1.2%	2,747	3.1%	0	0.0%
Allotments	7,018	14.5%	923	1.0%	0	0.0%
Cemeteries	8,955	18.4%	3,208	3.6%	0	0.0%
Sea	1,489	3.1%	3,105	3.5%	861	100.0%
Estuary	370	0.8%	2,292	2.6%	0	0.0%
River	9,343	19.2%	54,113	60.7%	0	0.0%
Lake	1,509	3.1%	14,397	16.1%	0	0.0%

Similar data on the presence of different forms of formal designation are provided in Table 4. Note that the category 'nature' includes numerous form of designation for nature protection including local and national nature reserves, Natura 2000 sites, Ramsar Sites, SSSIs and Ancient Woodlands.

Table 4: Designations present at recreation sites

Designation	Parks		Paths		Beaches	
	Number	Percent	Number	Percent	Number	Percent
National Park	2,389	4.9%	11,245	12.6%	57	6.6%
AONB	3,689	7.6%	17,172	19.2%	272	31.6%
Heritage Coast	448	0.9%	2,202	2.5%	222	25.8%
National Trail	672	1.4%	6,083	6.8%	330	38.3%
Historic Park	1,420	2.9%	6,431	7.2%	43	5.0%
Millennium Green	445	0.9%	163	0.2%	5	0.6%
Nature	5,059	10.4%	22,911	25.7%	650	75.5%
No Designation	32,384	66.7%	33,364	37.4%	90	10.5%

Table 5 provides details of the presence of different points of interest at recreational sites.

Table 5: Points of Interest present at recreation sites

Designation	Parks		Paths		Beaches	
	Number	Percent	Number	Percent	Number	Percent
Archaeological Feature	660	1.4%	8,696	9.7%	0	0.0%
Historic Building	654	1.3%	4,344	4.9%	0	0.0%
Scenic Feature	371	0.8%	4,735	5.3%	0	0.0%
Playground	8,700	17.9%	3,194	3.6%	0	0.0%
Viewpoint	511	1.1%	18,984	21.3%	0	0.0%
No Points of Interest	38,228	78.7%	70,228	78.7%	861	100.0%

3.6 Habitat Qualities

A key objective of the ORVal extension project was to provide a richer characterisation of the quality of certain habitats present at recreational sites and determine whether those quality characteristics could be seen to make a significant difference in recreational choice behaviour. One of the difficulties faced in extending the characterisation of habitat quality was that very few habitat quality datasets exist for the entirety of England and Wales. Accordingly, the choice of characteristics developed for the ORVal extension analysis tend to reflect data availability rather than necessarily capturing those aspect of environmental quality thought most likely to impact on recreation.

Woodland Type

The Forestry Commission's National Forest Inventory (NFI) provides an annual appraisal of woodland extent, composition and condition in Great Britain. The spatially referenced data categorises areas of woodland into the various categories that, for the purposes of the modelling exercise were reduced to the following three groups:

- Conifer (containing conifer and mixed mainly conifer categories)
- Broadleaved (containing broadleaf and mixed mainly broadleaf categories)
- Felled & Young Trees (containing the felled and young trees categories).

The original ORVal model contained a variable identifying the (natural log) of the area of a recreation site in woodland. In the extension model, additional variables were included indicating the percentage of that woodland area that was in each of the three tree type categories.

Wildlife Friendly Farming

The original ORVal model included a variable indicating the area of agricultural land traversed by a path used for recreational activity. To further characterise that agricultural land, DEFRA's spatially referenced records of farmland under Higher Level Stewardship schemes was used to identify the area of land under some form of government supported wildlife friendly farming.

Beach Littoral Sediment

Using data from the http://britishbeaches.info website, beach recreational sites were classified according to their dominant littoral sediment into one of the following four categories;

- Sand
- Shingle
- Sand & shingle
- Rocky or Harbour

Dummy variables indicating membership of these categories were included to characterise beaches.

Bathing Water Quality

The Environment Agency in England and Natural Resources Wales measure water quality at designated bathing water sites. Annual ratings are released that classify bathing water quality on a scale from poor to excellent. For the purposes of the ORVal model, each beach was identified as;

- High quality (containing sites classified as excellent or good)
- Low quality (containing sites classified as sufficient or poor)

Dummy variables indicating bathing water quality were included in the model for each beach.

River Ecological Status

As part of Water Framework Directive reporting requirements, the Environment Agency and Natural Resources Wales collect data on river water quality releasing that data on a six year cycle (Cycle 1 – 2009 and Cycle 2 – 2015). Using the smallest spatial scale at which that data is released (water body) the WFD ecological status categorisation of rivers passing through or by recreational sites was identified. For the purposes of the ORVal model, river ecological status classifications were combined into two groups

- High quality (containing sites classified as high or good)
- Low quality (containing sites classified as moderate, poor or bad)

Dummy variables indicating the ecological status of any river passing through or by a recreational site were included in the model

3.7 Weather

The Met Office MIDAS data archive provides daily weather details for meteorological stations across the UK over the time span of the MENE dataset. Of the various measures recorded in that data, those indicating daily rainfall and daytime temperature were chosen as potentially influencing recreational decisions. Attributing weather measurements to each respondent for each day of their seven day diary required some significant data processing. First, for each respondent the set of weather stations nearest to their home location were identified. Second, for each day of their seven day diary of recreational activity, a weighted average of the weather measures recorded at those neighbouring weather stations was used to estimate the weather at their home location. Rather than a simple linear interpolation the weighting variables were calculated using a natural neighbour interpolation method made available through the CGAL C++ library of spatial data processing routines.

4. Data Processing

4.1 Basic Observation Classification

The basic unit of observation in our data is a respondent-day; that is to say, the choice of outdoor recreation activity made by a respondent on a particular day. Since each respondent in the MENE data set provides information on their recreation activity over 7 days, each respondent contributes 7 different observations to the data.

For each of those observations the MENE data reveals whether or not the respondent took an outdoor recreation trip on that day. For the observation constituting the focus trip (the randomly selected trip for which detailed information is selected), MENE also provides information from which we might identify the recreation site visited and how they travelled to that site. Accordingly, at a basic level we can classify observations into one of three groups;

- No trip taken
- Trip taken to unidentified site
- Trip taken to identified site

For reasons not reported in the MENE documentation, the home location of some respondents is not recorded. Since the ORVal model requires information on how far different recreation sites are from a respondent's home, we were forced to drop these 8,326 observations from the dataset. In a similar vein, a key new dimension of the new ORVal model concerns choice over mode of transport. A further 636 observations failed to record the mode of transport used on the focus trip. Under the assumption that such miscoding of the data was random, those observations were also dropped from the sample.

Following the removal of observations from the dataset, demographic weights for the remaining sample were recalculated using the 'Anesrake' package for the R statistical software. The new weights ensured that the reduced sample could still be reweighted so as to be representative of the English population.

The remaining 316,959 observations returned complete and valid data on recreation activity. All the same, a number of reclassifications were necessary. For a start, for some observations the focus trip was reported as starting out from a location that was not the respondent's home. One possibility for explaining such responses is that the respondent was not at their home for the period covered by the survey perhaps staying with friends or on holiday. Since, the ORVal model focuses exclusively on day trips (as opposed to overnight trips) for the purposes of outdoor recreation, the 5,164 observations, rather than dropping those observations we reclassified them as being observations for which the respondent chosen the 'outside option', that is to say, had chosen something other than to take a recreational *day trip* to a greenspace.

In addition, the data recorded 608 respondents that had taken a trip to a village. Since visiting a village does not specifically imply engagement with greenspace, those observations were reclassified as choosing the outside option. In a similar vein 4,506 respondent's described the trip they had taken as being for the purposes of cycling or viewing greenspace from their car. Since both these transport based forms of greenspace recreation do not allow us to tie down enjoyment of the activity to some particular location, we again classify these observations as choosing the outside option. While, the 1,213 observations of trips to play golf do involve engaging with the environment in a particular location, that form of leisure activity is different from those to open access greenspace recreation sites insomuch as playing golf usually involves the payment of an entry fee (green fee). Accordingly, we classify these trips as constituting choice of the outside option.

Finally, 34,677 observations in the MENE dataset recorded a trip to open access greenspace but the record did not report the location of the greenspace visited. These observations were taken as being choice of the 'inside option' (that is to say, of taking a recreation trip) but that the data did not reveal the exact site chosen.

Following reclassification the dataset consisted of 202,121 observations where respondents had not taken a trip to greenspace over the course of the week previous to the interview, 34,677 observations where the focus trip was to an open access greenspace but we were unable to ascertain which particular greenspace and 80,886 where the focus trip was to an identified recreation site.

4.2 Choice of Transport Mode

One of the shortcomings of the original ORVal model was that it made the assumption that all outdoor recreation trips are made by private car. As such, the time and expenditure costs for all trips are based on those associated with driving. As shown in Table 6, however, only some 38% of the trips observed in the MENE dataset are taken by car. The majority, 52%, of trips are made on foot. Clearly assuming that visitors regard the costs of a visit made on foot as the same as that made in a car is unlikely to be correct. Walking is likely to be seen as cheaper, insomuch as there is no direct expenditure on market goods (fuel), equally walking is more costly in time, such that the direction of any bias is not easy to deduce.

In re-estimating the ORVal model, transport mode has been included as a second dimension of visit choice. Accordingly, each respondent is assumed not only to have a choice over travelling to every site in their choice set by car, but also through walking and, potentially, by each other means of transport. In practical terms, that causes some difficulties. The choice set of possible locations to visit for each respondent is very large, in theory as large as the 138,617 greenspace sites identified as available in England and Wales. Each time we add an alternative transport mode, we effectively add a further 138,617 options to the choice set; that is to say, the options of going to each of those sites by an alternative mode. Including those options for each of the 12 categories of transport model recorded in the MENE data would rapidly make the size of the estimation task quickly becomes unfeasible. In developing the new ORVal model, two possible aggregations over transport mode were considered (see final columns of Table 6):

- Four-Mode Classification: Walk, Car, Public, Bicycle
- Two-Mode Classification: Walk, Car

Table 6: Transport mode used for focus trips from MENE 2009-15

MENE Transport Mode	Number	Percent	4 Category	2 Category
On foot/ walking	58,286	51.85%	Walk	Walk
Car/van	43,127	38.36%	Car	Car
Bicycle/ mountain bike	3,521	3.13%	Bicycle	Walk
Public bus or coach (scheduled service)	3,362	2.99%	Public	Car
Train (includes tube/underground)	2,340	2.08%	Public	Car
Coach trip/ private coach	596	0.53%	Public	Car
Other	384	0.34%	-	-
Motorcycle/ scooter	274	0.24%	Car	Car
Taxi	224	0.20%	Car	Car
Wheelchair/mobility scooter	137	0.12%	Walk	Walk
On horseback	98	0.09%	-	-
Boat (sail or motor)	73	0.06%	-	-

The advantage of the four-mode classification is that it more faithfully replicates the key differences in modes chosen. The advantage of the two-class classification is that it captures the key distinction between motorised and non-motorised transport with the minimum of categories. Ultimately, and after discussion with an external referee, we opted for the two-mode classification, leaving open the possibility of extending the model to other modes at some future date.

4.3 Destination Matching

A first step in bringing together the MENE dataset and the ORVal Greenspace map requires matching the geocoded destinations for focus visits with the recreation sites identified in the greenspace map. Using destination details provided by the respondent (but not recorded in the released data) the survey administrators managed to attribute a six digit BNG reference to some 80% of the focus visits recorded in the survey (Natural England 2015).

The procedure for matching the MENE destination locations with the ORVal Greenspace map focused on the 80,886 respondents that had taken an outdoor recreation trip to a greenspace during the week in which they were interviewed and for which a valid geolocation was provided in the MENE data. Note that our analysis does not address the complicating issue of multi-site trips; the MENE data fails to record the information that would allow a proper characterisation of such trips. Accordingly, each trip is assumed to be solely for the purpose of visiting the site identified by the MENE destination location.

A scoring procedure was developed to facilitate the process of matching MENE destination locations with the ORVal Greenspace Map. In short, for each focus visit all recreation sites within 2.5km of the destination location recorded in MENE were identified. Details of each of those sites were then compared to information provided by the MENE survey and scored according to how well they tallied with details of the actual site visited in terms of their location, environs, site type and landcovers. The weights used to determine scores in the matching procedure were calibrated through examining how well the matching algorithm performed with a training data set where the actual destination could be readily determined from the data provided in MENE. Details of the matching algorithm and the weights used in the procedure can be found in Appendix I.

The matching algorithm took approximately 10 hours to run and identified a best guess as to the site on the ORVal Greenspace Map that was considered the mostly likely destination of each focus visit. As shown in Figure 2, where the score for each observation has been plotted in ascending order of score, matching scores varied across the range of 0 to 128.

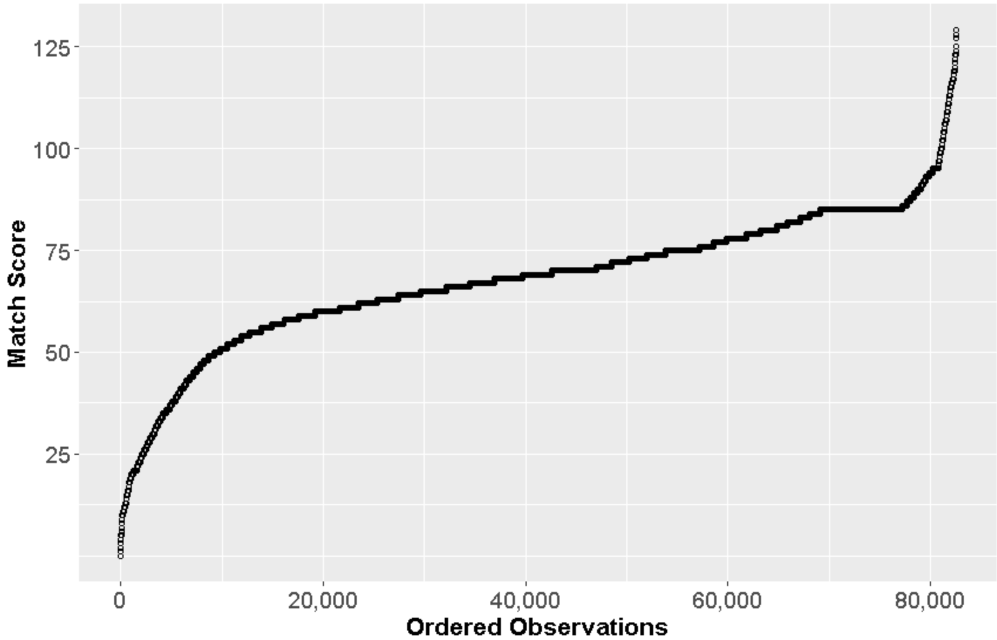

Figure 2: Matching scores for each focus trip plotted in ascending order of score

Through inspection of Figure 2, the change in slope of the data around the 50 pts mark was identified as a natural point to split the data. For the roughly 10% of focus trip observations below that threshold, the level of match was deemed too low to believe we had identified the actual site visited. Accordingly, those observations were reclassified into the "trip to unidentified site" category.

4.4 Respondent Sampling

Even with unusable observations removed the remaining dataset contained 317,684 observations. Since the estimation procedure to be used in the analysis (to be described subsequently) was relatively complex, it was decided to further reduce the dataset by drawing a smaller sample from those observations.

To ensure the richness of the data was maintained in that sampling procedure, a process of stratified random sampling was adopted where strata were defined by a respondent's choice of recreation activity on the focus trip. Accordingly, strata were defined as; (i) observations where no trip was taken; (ii) observations where a trip was taken to an unidentified site and (iii) a further 42 strata defined for observations where trips were taken to sites of different types. Those types were based on a classification of sites defined along four dimensions;

- Type: Beach, Park, Path, Woods, Nature, Allotment, Cemetery or Country Park
- Dominant Land Cover/Use: Woods, Sea Water, Fresh Water, Managed Grass, Agriculture, Natural Grass, Wetlands, Moors & Heath, Allotment or Cemetery
- Dominant Designation: National Park, AONB, Heritage Coast, Nature (including local and national nature reserves, Natura 2000 and Ramsar sites, ancient woodland), National Trail, Forestry Commission, Millennium & Doorstep Green, Historic Park or Country Park
- Points of Interest: Whether or not the site had archaeological remains, a historic building or a scenic feature or a viewpoint.

Applying this this four-dimensional classification scheme to the 138,000 recreation sites in the ORVal Greenspace Map resulted in 493 unique classes of sites. Naturally some of those classes contained very few sites such that classes were further aggregated so as to ensure that the MENE data set contained at least 100 focus visits to sites in each group. Definitions of the 42 groups identified through this procedure are shown in Table 7 and represent the choice-based strata used for sampling. The third column of Table 7 shows the number of observations with a focus trip to sites in each strata.

In order to establish our reduced sample, we used a stratified random sampling method in which we randomly sampled a fixed proportion of observations from each strata. To ensure representation of less commonly taken trips in the sample, the proportion taken for each strata was increasing in the rarity of visits. So a 20% sample was taken from strata with greater than 10,000 observed visits, a 30% sample for strata with between 7,500 and 10,000 visits, a 40% sample from strata with between 4,000 and 7,500 visits, a 50% sample from strata with between 3,000 and 4,000 visits, a 60% sample from strata with between 2,000 and 3,000 visits and a 75% sample for strata with less than 2,000 visits. The sampling probabilities and number of sampled observations are shown in columns 4 and 5 of Table 7.

Note that to correct for sampling bias in the MENE survey and to ensure representativeness of the sample, observations were drawn from strata in proportion to their demographic weights; that procedure increased the likelihood of drawing respondents with under-represented demographic profiles and decreased the likelihood of drawing respondents with over-represented demographic profiles.

A sampling weight was determined for each strata (described as the WESML weight in Table 7) that would late be used in estimation to correct for the choice-based sampling in the selection of observations (see Section 5). That weight indicates the ratio of the likelihood of a respondent drawn at random from the population having a focus trip to a site in a certain strata, to the likelihood of such an observation being in the sample. The population likelihood was estimated using the demographic weights for the full sample (see section 3.1) and the sample likelihood calculated from the numbers drawn from each strata;

Table 7: Choice-based sampling scheme and weights for (WESML) estimation

	Description	Num Obs	Sample Probability	Num Sample	WESML Weight
1	All Beaches	4,260	0.4	1,704	0.540
2	All Cemeteries	2,962	0.6	1,777	0.349
3	All Allotments	585	0.75	439	0.277
5	All Country Park	4,755	0.4	1,902	0.538
6	Path, Agriculture	4,227	0.4	1,691	0.565
7	Park, Managed Grass	20,794	0.2	4,159	1.002
8	Path, Managed Grass	1,705	0.75	1,279	0.297
9	Path, Agriculture, Nature	1,637	0.75	1,228	0.304
10	Path, Managed Grass, Nature	840	0.75	630	0.302
11	Path, Agriculture, AONB	445	0.75	334	0.305
12	Path, Managed Grass, AONB	389	0.75	292	0.306
13	Woods, Woods	1,369	0.75	1,027	0.274
14	Path, Woods, Nature	806	0.75	604	0.299
15	Path, Managed Grass, Natl Park	294	0.75	220	0.306
16	Woods, Woods, Nature	1,339	0.75	1,004	0.299
17	Path, Woods, AONB	373	0.75	280	0.308
18	Path, Agriculture, AONB, POI	338	0.75	254	0.301
19	Path, Agriculture, Nature, POI	438	0.75	328	0.301
20	Path, Agriculture, POI	544	0.75	408	0.302
21	Path, Woods	468	0.75	351	0.289
22	Path, Woods, Nature, POI	392	0.75	294	0.299
23	Park, Woods	3,582	0.5	1,791	0.397
24	Path, Fresh Water	1,173	0.75	880	0.286
25	Path, Moors & Heath, Nature, POI	289	0.75	217	0.319
26	Nature, Woods, Nature	1,635	0.75	1,226	0.280
27	Path, Managed Grass, NP, POI	197	0.75	148	0.319
28	Path, Managed Grass, AONB, POI	285	0.75	214	0.295
29	Path Woods National Park	145	0.75	109	0.310
30	Woods, Woods, AONB	360	0.75	270	0.309
31	Path, Managed Grass, Nature, POI	313	0.75	235	0.286
32	Other Fresh Water	3,996	0.5	1,998	0.421
33	Other Sea Water	1,575	0.75	1,181	0.280
34	Other Moors & Heath	577	0.75	433	0.308
35	All Wetlands	210	0.75	158	0.291
37	All National Trail	656	0.75	492	0.282
38	All National Park	1,026	0.75	770	0.305
39	Others No Designation	2,184	0.6	1,310	0.355
40	Other Nature Designation	3,036	0.5	1,518	0.424
41	All Historic Designation	8,744	0.3	2,623	0.696
42	All Heritage Coast	196	0.75	147	0.296
43	All Millennium & Doorstep Greens	471	0.75	353	0.267
44	All Forestry Commission	346	0.75	260	0.308
45	Other AONB Designation	930	0.75	698	0.306

$$w_{s(j_i)} = \frac{Q_{s(j_i^*)}}{H_{s(j_i^*)}} = \frac{Prop\ of\ Pop\ in\ s}{Prop\ of\ Sample\ in\ s} \quad (i = 1, 2, \ldots N) \tag{1}$$

where j_i^* indicates the site chosen by respondent i, $s(j_i^*)$ identifies the sampling strata for that site and $Q_{s(j_i^*)}$ and $H_{s(j_i^*)}$ are defined as shown in Equation (1).

The final estimation dataset comprised a sample of 64,383 observations, where each observation identified recreation behaviour over 7 consecutive days.

4.5 Choice Set Sampling

With the sample of observations to be used in estimation established, the next step in developing the dataset was to define the choice set for each respondent in the sample. In the case of the ORVal extension model, the choice set was two dimensional, comprising a choice over transport mode (walk or vehicle) and recreation site to visit. The other option in the choice set, of course, is the outside option; the choice not to take a recreational trip to outdoor greenspace on a particular day.

The issue of how to establish choice sets remains an open question in the literature; for a recent review see Thiene, Swait et al. (2017). In this research we assumed that each respondent's choice set consists of all recreation sites in England and Wales though, of course, many would be too distant from a respondent's house to ever compete with more proximate recreation sites offering similar experiences. Since the ORVal Greenspace Map identifies 138,617 sites, including each of these explicitly in the choice set of each observation for both modes of transport would result in an intractably large estimation dataset. Accordingly, we adopt a form of importance sampling in order to select a sample of sites for each observation and for each mode with which to model the full choice set.

To select the choice set sampled for a transport mode for each respondent, we wanted to ensure that the selected sites included;

(i) a diverse range of different outdoor greenspaces and
(ii) sites that were likely to be important possible recreation locations for that respondent.

To achieve (i), we again used a stratified sampling approach. Sites were categorised into 19 different strata according to their type and dominant landcover. The descriptions of those strata definitions are provided in the second column of Table 8.

As shown in the final column of Table 8, a sampling scheme was devised in which the number of sites sampled from a category to be included in an individual's choice set was selected according to the number of sites in each category type. So a category containing more than 10,000 sites (e.g. paths through agricultural land) was sampled 8 times, a category with greater than 1,000 but less than 10,000 sites was sampled 4 times and a category with less than 1,000 sites was sampled twice. Where the respondent had taken a trip to particular greenspace, that greenspace was included in their choice set and one less alternative sampled from the category corresponding to the chosen site. That sampling strategy lead to a selection of 83 sites, with the sampling process being repeated for both modes of transport. In other words, we selected 83 sites with which to represent trips to sites that might be taken in a vehicle and a separate independent sample of 83 sites to represent sites that might be visited on foot. A final category of not taking a trip at all was added to the choice set giving a (sampled) choice set size of 163 options.

Table 8: Choice set sampling scheme

Category	Description	Num Sites	Num Sampled
0	No Trip	0	1
1	Beaches	861	2
2	Cemeteries	9,013	4
3	Allotments	7,063	4
4	Parks mostly woods	12,481	8
5	Parks mostly wetland	140	2
6	Parks by sea	245	2
7	Parks mostly natural grass	1,184	2
8	Parks mostly moorland	216	2
9	Parks mostly managed grass	16,996	8
10	Parks by fresh water	1,191	4
11	Parks mostly agricultural	19	4
12	Paths mostly woods	11,538	8
13	Paths mostly wetland	169	2
14	Paths by sea	1,110	2
15	Paths mostly natural grass	3,267	4
16	Paths mostly moorland	3,209	4
17	Paths mostly managed grass	25,514	8
18	Paths by fresh water	3,893	4
19	Paths mostly agricultural	40,512	8
	Total:	138,621	83

To ensure (ii) (that is, that 'important' sites were selected from each strata for each mode) we first calculated the straight line distance between the centroid of each site and that respondent's home location (as identified by their 6 digit postcode). Then for each strata we selected random samples of sites with sampling weight proportional to the inverse of distance for the car mode options and proportional to inverse distance squared for the walking mode options.[1]

[1] To provide a brief intuition as to the functioning of these choice set sampling weights, in estimation we are going to need to calculate a sum across all the sites in a respondent's real choice set; roughly speaking adding up the utility the respondent might have got if they had chosen to visit each site. So imagine that there were four sites in the real choice set and let us label the utility from visiting each of those sites as u_1, u_2, u_3 and u_4. Our best guess is that $u_1 > u_2 > u_3 = u_4$ in the ratio 4:2:1:1. Now imagine we wanted to estimate the sum $u_1 + u_2 + u_3 + u_4$ but could only base our guess on that sum through drawing a sample of one observation. Given our best guess of the relative sizes of the four utilities, we could use importance sampling which means we would sample u_1 with probability $4/8 = 1/2$, u_2 with probability $2/8 = 1/4$ and u_3 and u_4 both with probability $1/8$. Now if we were to draw site 1 as the single observation in our sample, the weight in (3) would be 1 divided by $1/2$ which is 2. So our best bet at the sum $u_1 + u_2 + u_3 + u_4$ given this single observation would be $2u_1$. Likewise if we were to draw site 2, our best estimate using our importance weights would be $8u_2$. If instead we were to draw a sample of two observations, there would be 2 chances of selecting any site into the choice set such that in calculating weights the denominator of (3) would be doubled. Say we drew

4.6 Travel and Time Cost Calculation

The final step in generating an estimation data set was the calculation of travel costs for each option in the choice set. For that purpose we used the detailed Integrated Transport Network (ITN) data set provided by the Ordnance Survey. As shown in Figure 3 the ITN network not only identifies the roads suitable for travelling by car, but also an associated paths network which allows calculation of walking distances through the combined road and path network.

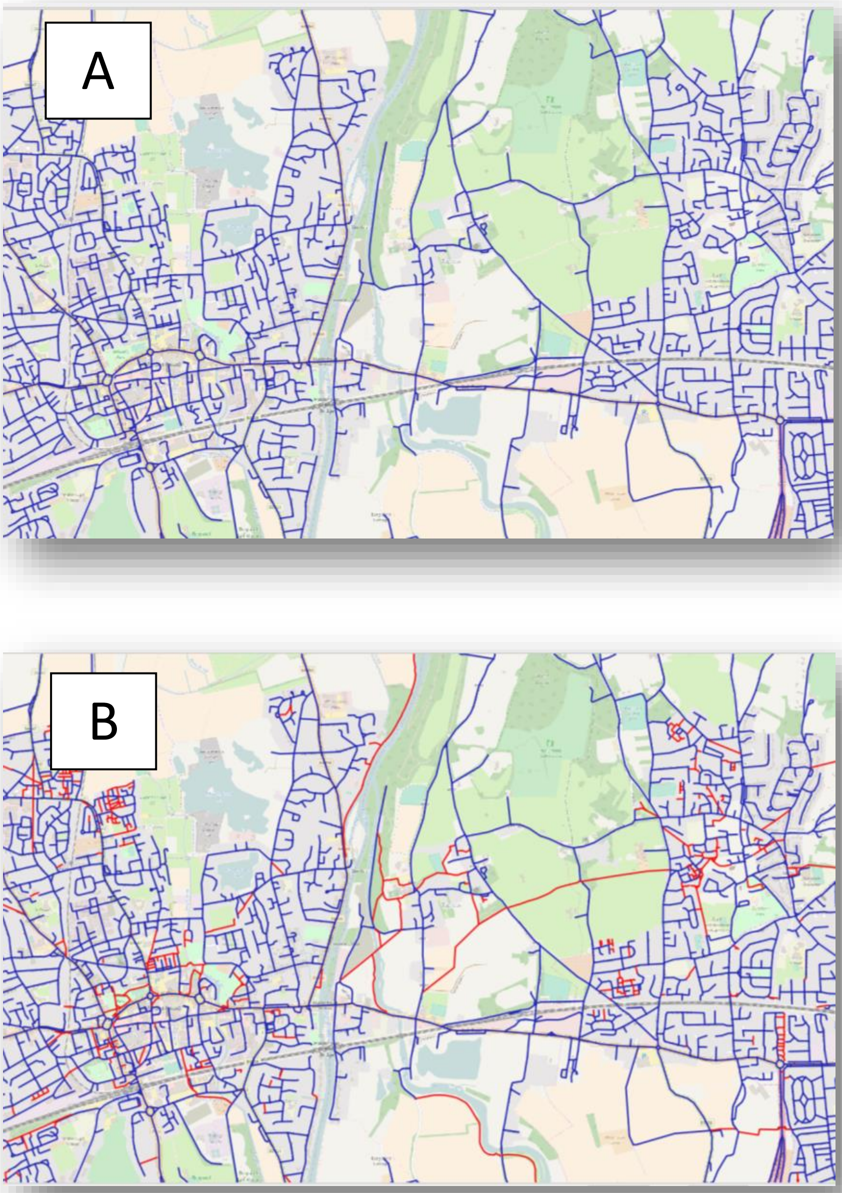

Figure 3: Detail of the OS Integrated Transport Network (A - roads; B - roads and paths)

---

sites 2 and 3. The weight for 2 would be $1/(2 \times 2/8) = 2$ and the weight for 3 would be $1/(2 \times 1/8) = 4$. Accordingly, given that sample of two observations selected through importance sampling our best estimate of the sum $u_1 + u_2 + u_3 + u_4$ would be $2u_2 + 4u_3$.

Our dataset consisted of a sample of 64,383 individuals for which driving times and costs must be calculated for trips to the 83 sites in the car mode choice set as well as walking times for trips to a further 83 sites in the walk mode choice set. Accordingly, our data demanded that we execute a total of 10,558,812 routing queries through the ITN network. The scale of that challenge was far in excess of the performance provided by standard routing software such as that available in the ArcGIS package. Accordingly, we turned to RoutingKIt, a highly efficient library of routing algorithms provided as C++ source code and applying the Contraction Hierarchy method of routing analysis (https://github.com/RoutingKit/RoutingKit). Having written the wrapper code to link RoutingKit to our data, we achieved speed ups of three orders of magnitude over that provided by ArcGIS's Network Analyst.

In running the routing queries we initialised the network with data on driving speeds along different categories of road and used RoutingKit to identify the fastest route from the home location to a site. For walking options we assumed a standard walking speed of 5kph. Using formulae provided by DFT (Department for Transport 2014), we took the length of time driving at different speeds along the route to calculate a fuel consumption for an average family car. Subsequently, we calculated a fuel cost by multiplying fuel consumption by the price of fuel at the time the trip was taken (taken as an average of diesel and unleaded prices from AA fuel price reports).

We converted driving times into a monetary cost using results provided in recent research for DfT on the value of travel time (Department for Transport 2015). Those values were £2.30 per hour for trips under 8km, £3.47 per hour for trips between 8km and 32km, £6.14 per hour for trips between 32km and 160km and £9.25 per hour for trips greater than 160km (see Table 7.18 of DfT report). A total monetary cost for driving to a site was taken by adding the time costs to the fuel costs for the return journey.

For walking options, that same DfT report indicates a value of travel time of 7.6 p per min or £4.58 per hour. While a linear opportunity cost of time in walking might hold true over reasonably short trips, for longer walking trips an escalating value of time, similar to that identified for driving time costs, is likely more realistic. Accordingly we transformed walking travel times to a cost of travel using the following function:

$$time\ cost = (\exp(hrs\ travel\ time) - 1) * £4.58$$

That specification ensures that the cost of travel time progressively escalates with increasing travel time with the marginal cost of time taking the value £4.58 for an hour long trip (e.g. half an hour there and back) and increasing exponentially thereafter. Since the exponential function resulted in some very large cost of travel time for those respondents whose choice sets included very distance sites, we censored travel time costs at a value of £675, an amount which approximates 5 hours walking time.

5. The Econometric Model

5.1 Econometric Specification

Our approach to estimating a recreational demand model adopts the long-established random utility framework first proposed by McFadden (1973). That framework characterises recreational decisions as discrete choices in which, on any particular choice occasion, an individual has the opportunity to visit one of an array of sites each offering different opportunities for outdoor recreational activities. In essence, the modelling approach seeks to establish the value of the recreational opportunities offered by sites by observing data recording which particular sites individuals chose to visit given the set of sites that they could have possibly visited.

More formally, our dataset records the outdoor recreational choices of a sample of individuals, indexed $i = 1, 2, \dots, N$, on each of series of days indexed $t = 1, 2, \dots, T$. That recreational choice concerns which greenspace to visit where greenspaces are indexed $j = 1, 2, \dots, J$ and by which mode of transport that greenspace is accessed, indexed $h = 1, 2, \dots, H$. In our application where we assume only walking or driving transport modes, $H = 2$. A final option concerns whether to undertake some other activity, an option indexed $j = 0$.

The choice as to which greenspace to visit depends on a number of factors, but two important considerations are the quality of the recreational experience offered by a site and the cost in time and money of visiting that site. That cost, of course, differs according to the choice of mode of transport. In our model, the quality of recreational experience offered by site j is determined by the vector of site characteristics $\boldsymbol{x_j}$ and the costs of making a trip to that site using a certain mode of transport, tc_{ijh}.

To construct our econometric model, we first need to posit a function which describes the utility an individual will enjoy if they decided to visit site j . In line with the vast majority of the literature we choose the simple linear approximation;

$$v_{ijht} = \alpha_j + \boldsymbol{x}_j\boldsymbol{\beta_1} + \gamma\left(I_{i,t} - tc_{ij}\right) \qquad (j = 1, 2, \dots, J \text{ and } \forall i, t, h) \tag{2}$$

where, $I_{i,t}$ is individual i's per period income, α_j is a site-specific utility element, $\boldsymbol{\beta_1}$ is the vector of coefficients describing the marginal utilities of site qualities and γ is the marginal utility of income.

Alternatively, an individual may choose not to make an outdoor recreational trip. We give that "no trip" option the index $j = 0$, and specify the utility from that option as;

$$v_{i0t} = \alpha_0 + \boldsymbol{z}_{i,t}\boldsymbol{\beta_0} \qquad (\forall i, t) \tag{3}$$

where $\boldsymbol{z}_{i,t}$ is a vector capturing characteristics of the time period (e.g. month of the year, day of the week) and of the individual (e.g. gender, age, socioeconomic segment) whose importance in determining participation in outdoor recreation is captured by the vector of coefficients $\boldsymbol{\beta_0}$, while α_0 is some constant utility associated with choosing not to take a trip to greenspace.

Adopting the familiar random utility framework, we develop our econometric specification from (2) and (3) by constructing the conditional indirect utility function;

$$u_{ijht} = v_{ijht} + \varepsilon_{ijht} \qquad (j = 0, 1, \dots, J \ and\ \forall i, h, t) \tag{4}$$

where ε_{ijmt} is an econometric error term introduced to capture the divergence between our model of utility $\left(v_{ijht}\right)$ and the individual's experienced utility $\left(u_{ijht}\right)$. Since the scale on which utility is measured is not known, we can make any arbitrary decision as to what quantity represent zero. For the purposes of this analysis we set $\alpha_j = \alpha = 0\ \forall j$. Given the very large number of sites in the analysis, this assumption amounts to relegating utility derived from idiosyncratic features of each parks to the error term.

To simplify notation, henceforth we drop the mode and time subscripts. Given that presentational simplification, we progress by assuming that individuals choose from the set of options $j = 0, 1, \dots, J$, selecting that option which gives them the highest utility. Accordingly, the probability of observing individual i choosing to visit site k can be written as;

$$\begin{aligned} P_{ik} &= Prob\left[u_{ik} > u_{ij} \ \ \forall j \neq k\right] \\ &= Prob\left[v_{ik} + \varepsilon_{ik} > v_{ij} + \varepsilon_{ij} \ \ \forall j \neq k\right] \\ &= Prob\left[v_{ik} - v_{ij} \ > \varepsilon_{ij} - \varepsilon_{ik} \ \ \forall j \neq k\right] \end{aligned} \tag{5}$$

Given v_{ik} and v_{ij} are, given parameters α_0, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are deterministic, the probability in (5) is determined by the assumptions made regarding the joint distribution of the error terms, $\varepsilon_i = \left[\varepsilon_{i0}, \varepsilon_{i1}, \dots, \varepsilon_{iJ}\right]$.

Perhaps the simplest assumption, and one used extensively in the choice modelling literature, is to assume that the error terms are drawn from the family of distributions described as Generalised Extreme Value (GEV) (McFadden 1978). In that case, the probability in (5) is given by;

$$P_{ij} = \frac{e^{v_{ij} + lnG_j}}{\sum_{k=1}^{J} e^{v_{ik} + lnG_k}} \tag{6}$$

Where the function $G(\cdot)$ follows from the particular assumptions made regarding the join distribution of the error terms and must conform to certain properties outlined by McFadden (1978). Also $G_j = \partial G / \partial e^{v_{ij}}$. The simplest form for GEV results from the assuming that;

$$G = \sum_{j} e^{v_{ij}} \tag{7}$$

Which, from (5), results in choice probabilities that define the familiar multinomial logit (MNL) model;

$$P_{ij} = \frac{e^{v_{ij}}}{\sum_{k=0}^{J} e^{v_{ik}}} \tag{8}$$

The great advantage of the MNL model is the simplicity of calculation of choice probabilities which greatly increases computational efficiency in estimating the model parameters. On the other hand the MNL model fails to allow for any form of correlation in the error terms of the different options or their observed attributes, an assumption that leads to predictions of somewhat implausible substitution patterns often referred to as *independence from irrelevant alternatives* (IIA) (McFadden, Tye et al. 1977). In effect, the IIA assumption does not allow for the expectation that the addition of a new option to the choice set will tend to reduce the probability of choosing options than have attributes more like that new option by a greater extent than it will options that are more dissimilar.

With the ORVal data, there are several dimensions of similarity over options that might determine the degree to which individuals regard them as substitutes. One similarity lies is in the type of recreational site. Respondents might well see recreational paths as being more similar to each other than other types; for example, beaches. A second possible dimension of similarity concerns the nature of habitats encountered while visiting a greenspace. Again we might expect that greenspaces with more similar landcovers, say woodlands, might be seen as being more similar to each other and hence close substitutes. Finally, respondents might regard walking options as being more similar to each other than they are to driving options. To capture these various possible dimensions of similarity, we defined a series of variables that identify each site's membership of various similarity groups;

- Type Category: Each recreation site was identified as being a;
  - Path
  - Park
  - Beach
  - Allotment
  - Cemeteries & Graveyards
- Habitat Category: We took the range of landcovers used to describe sites (see Table 3) and organised those into 10 broad categories;
  - Woods
  - Salt Water
  - Fresh Water
  - Natural Grass
  - Managed Grass
  - Agriculture
  - Wetlands
  - Moors & Heath
  - Allotments
  - Cemeteries & Graveyards

  Each site was allocated to a habitat category according to the landcover which constituted the largest proportion of that site's area.
- Transport Mode Category: Here driving and walking options are identified as being separate categories.

The possibility that options in the same category are considered closer substitutes can be handled through an alternative specification of the GEV model where the G function is defined as;

$$G = \sum_{m=0}^{M} \left(\sum_{j \in B_m} e^{\mu_m v_{ij}} \right)^{1/\mu_m} \tag{9}$$

Here the sites in each distinct category group form the set B_m and $m = 1, 2, \ldots, M$ indexes the different categories. Notice we have added an additional group, $m = 0$, which has the single member consisting of the option not to take a recreational trip. Notice also, the group-specific parameters, μ_m, which allow for the fact that sites in a category may be similar to each other in some unobserved way. As shown by McFadden (1978), this similarity parameter should vary on the range from 1 to ∞. When μ_m is large then individuals regard the sites in group m as very similar and hence treat them as close substitutes. In contrast when the $\mu_m = 1$ the sites in the group are considered no more similar to each other then they are to any other site; indeed if $\mu_m = 1$ for all m (9) reduces to (7) and we are back at the MNL model. Replacing (9) in (6) results in the specification of a GEV known as the nested multinomial logit model (NMNL) which is thoroughly reviewed and described in Morey (1999).

Constraining similarity between sites to be dictated by the dominant landcover ignores the fact that each site is actually a mosaic of landcovers such that each site may share similarities with a variety of groups. Accordingly we develop one further categorisation variable:

- Habitat Proportion: Using the same groups as used to identify habitat categories, a series of variables capturing the proportion of each recreation site under different habitats. Clearly, with this categorisation each site can belong to a number of categories to different degrees.

A specification of $G(\cdot)$ that accommodates the possibility of membership across multiple categories is given by;

$$G = \sum_{m=0}^{M} \left(\sum_{j=0}^{J} \alpha_{jm} e^{\mu_m v_{ij}} \right)^{1/\mu_m} \tag{10}$$

(10) differs from (9) with regards to the parameters α_{jm} which dictate the 'share' of site j that should be apportioned to similarity group m, for example, α_{jm} could capture the proportion of site j's land area that is of landcover m. With this specification, therefore, a site is seen as similar to other sites with which it shares landcovers but more similar to sites with which it has more landcover in common. Replacing (10) in (6) results in a specification of a GEV model known as the *cross nested logit model* (CNMNL) first proposed by Ben-Akiva and Bierlaire (1999) and reviewed in detail by Bierlaire (2006).

Compared to other possible GEV specifications, the CNMNL admits rich patterns of substitution between greenspaces that reflect the similarities in environmental experience offered by the different sites. From (6) we see that the mathematical form of the CNMNL choice probability, while more complex than the MNL model, remains reasonably tractable.

The programme of modelling undertaken in the ORVal extension project involved estimation and comparison of a series of GEV models:

- Multinomial Logit (MNL) model
- Nested Multinomial Logit (NMNL) model with nests defined by (a) type categories, (b) habitat categories and (c) transport mode categories.
- Cross Nested Multinomial Logit (CNMNL) model with nests defined by (a) habitat proportions, (b) landcover proportions and type category and (c) habitat proportions, type categories and transport mode categories. In the latter models assumptions need to be made regarding what 'share' of an observation to attribute to the type and mode category.

In passing we note that the partial derivative in (6) for the CNMNL takes the form;

$$G_j = \frac{\partial G}{\partial e^{v_{ij}}} = \sum_{m=0}^{M} \alpha_{jm} e^{v_{ijt}(\mu_m - 1)} \left(\sum_{j=0}^{J} \alpha_{jm} e^{v_{ij}\mu_m} \right)^{1/\mu_m - 1} \tag{11}$$

such that the choice probabilities can be written as;

$$P_{ij} = \frac{e^{v_j + \ln G_j\left(\sum_{q=0}^{J} \alpha_{qm} e^{v_q \mu_m}\right)}}{\sum_{k=0}^{J} e^{v_k + \ln G_k\left(\sum_{q=0}^{J} \alpha_{qm} e^{v_q \mu_m}\right)}} \tag{12}$$

Where the notation $G_j\left(\sum_{h=0}^{J} \alpha_{hm} e^{v_h \mu_m}\right)$ is included to make explicit the fact that the partial derivative (11) is a function of a sum across all greenspaces in the choice set.

One thing to note about the form of CNMNL model shown in (12) and adopted in the ORVal Greenspace Model is that it makes no accommodation for the fact that our data contains observations of the same individual making choices across multiple choice occasions.[2]

Given data on the recreational choices of the N individuals in T time periods, it follows from (12) that the log of the likelihood of observing those choices is;

$$\ln L(\alpha_0, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \gamma, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=0}^{J} Y_{ijt} \ln P_{ijt} \tag{13}$$

Where Y_{ij} is a dummy variable which takes the value 1 if individual i chose recreational option j, or zero otherwise and $\boldsymbol{\mu}$ is the vector of similarity parameters. The parameters of the model can be estimated using maximum likelihood methods by optimising (13) with respect to the parameters $\alpha_0, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \gamma$ and $\boldsymbol{\mu}$.[3]

Given the size of the dataset a significant programme of code development was undertaken to ensure the models could be estimated in a reasonable time. In particular, the models were written in C++ with the compiled estimating code achieving speed ups of some 30 to 50 times code written in a standard econometric matrix programming language.

5.2 Welfare Estimation

As shown in equation (5), an important feature of GEV models is that they are firmly based on a theory of random utility maximisation. Indeed, provided empirical estimation of the model results in $\mu_m \geq 1$ $(m = 0,1, \ldots, M)$ in the NMNL and CNMNL, then the model is globally consistent with that theory (Kling and Herriges 1995).

One useful property of GEV models that follows from that fact, is that there exists a simple closed-form expression for the expectation of the maximum utility a respondent might expect to derive from being able to choose an option from their choice set. In the case of the CNMNL model that expression amounts to;

---

[2] Note that subsequently we employ clustered robust standard errors to account for the lower information content provided by repeated responses from the same individual.

[3] Under two circumstances the MENE data records that the respondent has taken a trip on that choice occasion but does not record which greenspace was the visited. First, for days in the respondent's week-long dairy record where a trip was taken but that choice occasion was not randomly selected as the focus trip. Second, where we were unable to identify the location of the focus trips (see Section 4.2). On those occasions all we know is that the respondent chose to take a trip to some greenspace or, put another way, decided not to choose the outside option indexed 0. Accordingly, under both those circumstances, we record the probability of the choice as $1 - P_{i0t}$

$$V_{it}(J') = \ln \sum_{m=1}^{M} \left(\sum_{j \in J'} \alpha_{jm} e^{v_{ijt}\mu_m} \right)^{1/\mu_m} + \lambda \tag{14}$$

where $V_{it}(J')$ is the expectation of maximum utility realised by individual i in time period t given the opportunity to choose from the choice set J', and λ is the Euler-Mascheroni constant (that takes a value of 0.5772 to 4 decimal places).[4]

It follows that the expected level of welfare change that an individual would experience if the nature of their choice set were to change, perhaps through the loss or gain of sites from the choice set and/or changing the qualities of sites (Small and Rosen 1981);

$$\Delta W = \frac{1}{\gamma}\left(V_{it}(J'') - V_{it}(J')\right) \tag{15}$$

where J' is the original choice set and J'' the changed choice set. In simple terms, equation (15) describes the analyst's best estimate of how an individuals' utility will change as a result of changes in the choice set with that quantity translated into money terms by dividing that utility change by the marginal utility of income, γ.

5.3 Econometric Corrections

The econometric model as defined by (13) fails to correct for a number of features of the data used in estimation of the model. IN the first instance, the specification in (13) assumes random sampling, where the data used in estimating the ORVal model were drawn using the choice-based sampling strategy described in Section 4.3. To correct the likelihood we use the weighted exogenous sampling maximum likelihood (WESML) estimator as follows;

$$ln\, L = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=0}^{J} Y_{ijt} w_{s(j_i^*)}\, ln\, P_{ijt} \tag{16}$$

where $w_{s(j_i^*)}$ is the choice-based sampling weight defined in (1). In effect the weight acts to correct the log likelihood function, decreasing the importance of observations that have been over-sampled in drawing a choice-based sample and increasing the importance of observations that have been under-sample. Manski and Lerman (1977) show that the WESML estimator provides consistent estimates of the model parameters.

A second issue with the ORVal model data set is the sampling of sites for inclusion in the choice set for each sample respondent. Recall from (6) that the GEV probability for choice option j takes the form of a relatively simple proportion relating the utility from a visit to site j to an aggregations fo the utilities of visits to all sites in the choice set. More specifically, in the CNMNL model the numerator of the probability is the exponentiated utility from a visit to site j plus a term that captures the degree of similarity of that site to others in the choice set, while the denominator is the

[4] The derivation of this formula arises from interpreting the conditional indirect utilities of each option (see equation (6)) as random variables and calculating the expected maximum of that set.

sum of exponentiated utilities plus similarity terms for the entire choice set. Clearly, when we use a sample of options in the choice set two errors arise in the choice probability. First, the ratio of numerator to denominator is biased since we fail to sum over the full choice set in the denominator. Second the similarity term is biased since it fails to aggregate over all sites similar to site j.

To address these biases, Guevara and Ben-Akiva (2013) propose a correction to the choice probability of the form;

$$P_{ij} = \frac{w_{ij}^{1} e^{v_{ij}+\ln\hat{G}_j\left(\sum_{q\in\tilde{J}_i^2} w_{iq}^2 \alpha_{iqm} e^{v_q \mu_m}\right)}}{\sum_{k\in\tilde{J}_i^1} w_{ik}^{1} e^{v_{ik}+\ln\hat{G}_k\left(\sum_{q\in\tilde{J}_i^2} w_{iq}^2 \alpha_{iqm} e^{v_q \mu_m}\right)}} \tag{17}$$

where the weights w_{ij}^2 $\left(j \in \tilde{J}_i^1\right)$ are calculated as per (1) to reflect the relatively probability of an alternative appearing as one of the options in the sampled choice set, $\tilde{J}_i^1$. Likewise w_{iq}^2 corrects the aggregation over sites that appears in the similarity terms. Accordingly, we denote these similarity terms by the functions $\hat{G}_j(\cdot)$ to make clear that the aggregation over choice sets used in their calculation is an estimate based on the choice set sampling weights, w_{iq}^2 $\left(q \in \tilde{J}_i^2\right)$. Notice that as per the recommendation of Guevara and Ben-Akiva (2013) we sample a second set of options to form the choice set used to calculate the similarity terms, $\tilde{J}_i^2$ and the weights w_{iq}^2 are calculated from this sample as per (1).

5.4 Covariate Choice

The final step in developing the ORVal recreation demand model is to determine the set of covariates that will be used to describe the participation choice, $\boldsymbol{z}_{i,t}$ $(i = 1, \ldots, N, t = 1, \ldots, T)$, and those to be used to describe the utility benefits of a trip to a site $\boldsymbol{x}_j (j = 1, \ldots, J)$.

In the model, the choice of whether to take a trip or not was made a function of three groups of variables; those that described the time when the trip was taken, those that described the location of residence of a respondent, and those that described a respondent's socio-demographic characteristics;

1. Time: We captured the time dimension through a set of dummy variables for the year (using 2009) as the base case, a set of dummy variables for month of the year (using December as the base case) and a set of dummy variables for day of the week (using Monday as the base case). As part of the ORVal extension project we also introduced a dummy variable identifying Bank Holidays.
2. Location: Location of residence was represented by a set of dummy variables coding for a respondent's Government Office Region (GOR) using the East Midlands as a base case. As part of the ORVal extension project a dummy variable was introduced to identify residents of urban areas from those in rural areas.
3. Sociodemographics: A key consideration in defining variables to describe a respondent's sociodemographic characteristics was the subsequent need to transfer the ORVal recreation demand model to predict the behaviour of all (adult) individuals in England and Wales. While the MENE survey collected numerous details of the sociodemographics of the survey respondents, our information for the wider population is limited to data provided by the

2011 census and collated at the level of Lower Super Output Area (LSOA). Accordingly, our selection of variables by which to describe the sociodemographics of respondents was restricted to those provided in both the MENE survey and the 2011 census. In particular, we defined dummy variables identifying age-gender groups, individuals with children, working status, and socioeconomic segment (using the six category – A, B, C1, C2, D, E - socioeconomic classification produced by the ONS). While ownership of a dog was not recorded in the census we included this as a covariate and used data from elsewhere to approximate that value in the transfer exercise. As part of the ORVal extension project, a set of variables indicating the ethnicity of the respondent were included in the model specification. Those categories identified White, from Black, from Asian, from Mixed ethnicity with other racial groups falling into an 'Other' category.

As shown in (2), we assume that the utility derived from visiting a greenspace comprises the trade-off between a cost and a benefit. The cost comes in the form of the time and travel expenses incurred in getting to and from that greenspace; the travel cost, tc_{ij}, calculated as explained in Section 4.5. The benefits, it is assumed, are derived from the various qualities of the greenspace. We capture those qualities through a series of sets of covariates;

4. Greenspace Type: To establish differences in utility offered by different broad categories of greenspace, we created a dummy variable set distinguishing paths from beaches, from country parks, from allotments from graves/cemeteries leaving other parks (see definition in Section 3.3) as the base case.
5. Size and Landcover Composition: The nature of the greenspace with which an individual interacts when visiting a site is captured in the ORVal greenspace model through a series of variables that record the natural log of the total area of the greenspace (in hectares), and the natural log of the areas of each landcover from which that greenspace is composed. With regards to the latter we identify the area (in hectares) of each park dominated by the following 13 landcover types;
    - Woods
    - Wood Pasture
    - Agriculture
    - Natural Grass
    - Moors
    - Coastal
    - Saltmarsh
    - Marsh & Fen
    - Recreational Grass
    - Sports Pitches
    - Gardens
    - Allotments
    - Cemeteries

   Each landcover is represented in the specification through three variables. First we include dummy variables that identify sites for which a landcover type is the dominant landcover. Subsequently, we take the natural log of the total area of the recreation site and multiply this by the proportion of that area under a landcover giving a measure. That variable is introduced into the model in two forms; one as a variable specific to path recreation sites and the other as a variable specific to park recreation sites.

In a similar vein, we identify three different types of water area to be found in recreation sites:

- Sea
- Estuary
- River
- Lake

Again we introduce each of these water variables into the model through three variables; a constant for sites whose dominant habitat is one of the water area types plus park and path specific variables identifying the proportion of the natural log of total water area for a site under each of the water types.

In addition to the quantities of different land and water covers, the specification includes a variable which describes the diversity of landcovers accessible from a greenspace calculated using Simpson's Index of Diversity. In particular, we calculate the proportion of a greenspace under each land cover type according to;

$$prop_{l,j} = \frac{area\ under\ landcover\ l\ at\ site\ j}{total\ area\ of\ site\ j} \quad (l = 1, \dots, L; j = 1, \dots, J) \tag{18}$$

We then calculate the diversity index as;

$$diversity_j = \frac{1}{\sum_l^L prop_{l,j}{}^2} \quad (j = 1, \dots, J) \tag{19}$$

Observe that the lowest possible value of the index is 1 where the greenspace has only one landcover but increases in the number of landcovers accessible at that site. For example, with two land covers the index can take a value in the range 1 to 2, where an index near 1 would indicate only a small part of the greenspace having the second landcover and an index of 2 would arise when the greenspace has equal areas of the two landcovers. Likewise, with three landcovers the index can take a value in the range 1 to 3 with the upper bound again identifying an equal split of area between the three landcovers.

6. Commonalities: One complication with the definition of greenspaces is in defining what constitutes an independent recreation site, for example in circumstances where the ORVal greenspace map identifies greenspaces that share common boundaries (though see section 4.15 of the ORVal Greenspace Map Report). Ignoring the fact that a greenspace borders another greenspace may understate its qualities since individuals visiting that site may also take advantage of the greenspace provided by the adjoining site. In that case, we might think that commonalities between the borders of greenspaces might indicate complementarities not otherwise captured in our model.

   For path sites, defined as access points to a path network, the issue of commonality is likely to act in the opposite direction. In this case we define the commonality to be the area of overlap in the path network accessible from a particular access point. Where multiple path sites access the same path network then those different sites are likely to represent close substitutes. Accordingly, we might expect that a path site with more commonalities (and hence more close substitutes) will receive fewer visits than for an identical path site with no commonalities.

The issue of commonalities has received attention in the transport literature concerned with route choice. In the ORVal model we adopt the proposal of Cascetta, Nuzzolo et al. (1996) where they define a variable that captures the degree of commonality for option according to;

$$CF_j = \ln \sum_k \left(\frac{L_{jk}}{L_j^{1/2} L_k^{1/2}} \right)^{\rho} \tag{20}$$

Where CF_j is the commonality factor for option j, L_{jk} is the extent of commonality between site j and site k, L_j is the total extent of site j, L_k is the total extent of site k and ρ is a parameter that we set to the value 1 in calculation of the commonality factors. In the case of parks, the extent of a site is taken to be its perimeter and the extent of commonality with another park is the extent of that perimeter that lies within 25m of that other park. In the case of paths, the extent of a site is taken to be the linearly decayed area of path accessible from a path access point, and the commonality with another access point is the extent of that area accessible from that other access point.

7. Designations: The ORVal greenspace map records a variety of special designations given to the different recreational sites. For the purposes of the ORVal model we assume that those designations may capture aspects of the environmental experience of visiting a greenspace that are not captured by descriptions of type or landcover. Accordingly we define a series of binary variables identifying sites with the following designations.
    - National Park
    - AONB
    - CROW
    - Heritage Coast
    - Historic Park
    - Millennium or Doorstep Green
    - Nature

    Note that the 'Nature' category includes designation as a local nature reserve, national nature reserve, a Natura2000 site, a RAMSAR site, SSSI and ancient woodlands.
    We include the designations in the model in proportions; that is to say, as the proportion of the recreation site under each designation.

8. Points of Interest: The final set of variables used to describe the quality of greenspaces are a set of binary variables identifying the presence of a series of possible points of interest;
    - Archaeology
    - Historic Building
    - Scenic Feature
    - Playground
    - Viewpoint

Since we suspected that the recreational experience associated with a path-type site may differ from that of a park-type site (see definitions in Section 3.3) we define separate sets of site quality variables for paths and parks; that is to say, we have one set of size & landcover, commonality, designation and points of interest variables defined for parks and another set for paths.

6. Results

6.1 Specification Testing

An extensive programme of specification testing was undertaken in order to inform on the most appropriate form for the ORVal recreation demand model. In the first instance that involved experimenting with the appropriate selection of covariates and the functional form of the indirect utility functions for participation in outdoor recreation (Equation 3) and mode and site choice (Equation 2). The functional forms arrived at through that investigation included 63 variables describing the participation choice and 86 describing site and mode choice.

Before describing in detail the parameter estimates on those covariates, we begin by considering the question of which form of GEV model best captures patterns of similarity between recreation options. Table 9 lists out various measures of model fit across a range of GEV specifications.

Table 9: Fit statistics for different model specifications

Model	In Sample	Out-of-Sample			
	Choice	Choice	Participation	Conditional Mode & Site Choice	Conditional Mode
Multinomial Logit:					
Constants:	-372,808.28	-31,832.80	-16,715.53	-15,125.76	-975.88
Constants & TC	-290,233.54	-25,966.39	-17,191.97	-8,782.75	-904.32
Full	-256,007.94	-23,536.09	-15,580.55	-7,962.36	-862.62
Nested Logit:					
Habitats	-254,102.87	-24,349.82	-16,466.77	-7,888.49	-853.95
Type	-254,219.93	-24,759.59	-16,828.10	-7,936.77	-854.64
Mode	-253,235.22	-36,346.49	-28,016.30	-8,332.83	-1,379.90
Cross-Nested Logit:					
Habitats	-253,750.81	-24,286.87	-16,405.29	-7,887.07	-851.26
Habitats & Type	-253,249.47	-24,743.72	-16,890.64	-7,858.26	-854.88
Habitats, Type & Mode	-252,160.33	-25,981.77	-18,174.67	-7,811.56	-874.92

<u>In-Sample Goodness of Fit</u>

The first column of Table 9 shows the in-sample fit of each model. What is reported here is the maximised value of the log likelihood, that is to say, a statistic showing the log of the probability that the model assigns to observing the set of choices made by the respondents in the estimation dataset. The larger the log likelihood (that is, the closer the statistic is to zero) the better job the model does at matching those observed choices.

The first model in Table 9 is a simple MNL model in which we include just a single constant to capture preferences for taking a trip as opposed to choosing the outside option, a single constant to capture preferences for walking over driving and a set of four constants to distinguish preferences for paths, beaches, cemeteries and allotments over parks. This is our minimal specification against which to compare more sophisticated specifications.

The second MNL model in Table 9, adds in just two more covariates; the travel costs associated with travelling to a site by car and the travel costs associated with walking. Notice the very substantial increase in log-likelihood score from -372,808 to -290, 233. Not surprisingly, recreational choices are very significantly determined by the costs of travelling to sites; we can explain the observed recreational choices significantly better when we account for those costs.

The final MNL model in Table 9 includes the full set of 149 preference function covariates. Again we observe a highly significant increase in the goodness of fit of the model to the estimating dataset. It appears that the variables that we include to characterise participation, mode and site choice significantly improve the model's ability to distinguish between the estimation sample's observed choices.

The remaining models in Table 9 use the full set of preference function covariates but instead examine different form of GEV model to explore specifications which better capture patterns of similarity between sites.

The nested multinomial logit (NMNL) models ascribe options to particular similarity groups; for example, the NMNL Habitats model groups options into an outside option group and 10 habitat groups ascribing each recreational site option to the group of their dominant landcover. Likewise the NMNL Type model groups recreational sites by type categories (paths, parks, beaches, allotments, cemeteries & graveyards) while the NMNL Mode model groups together the walking options for each recreational site and the driving options for each recreational site. Without presenting the results in detail, the similarity parameters (μ_m from Equation 9) for each of these models are greater than 1 thus conforming to expectations from maximum utility theory and the majority are statistically significantly greater than 1.

Notice that the NMNL models each deliver an in-sample goodness of fit superior to the MNL model; likelihood ratio statistics show these to be highly significant at greater than the 99.9% confidence level. Notably accounting for similarity between modes leads to the highest log likelihood amongst the NMNL models of -253,235. Accordingly, the models suggest that accounting for patterns of similarity between options are important in explaining the choices of the estimation sample with an increased tendency to substitute within travel mode type providing the most significant increase in fit of the model.

The last three rows of Table 9 report details of Cross-Nested Multinomial Logit (CNMNL) specifications which allow for cross-nesting, that is to say, allow options to be part of more than one similarity group. Each option's membership of the different similarity groups is determined by the α_{jm} parameters of Equation 10. Those parameters are proportions such that summing α_{jm} across the different groups (m) for an option (j) returns a value of 1. In the CNMNL habitat model, for example, the α_{jm} are calculated as the proportion of the recreation site's area in the different habitat groups.

The second CNMNL model introduces the site type categories alongside the habitat proportions. In the model reported in Table 9, we impose the assumption that half of an option's group membership is determined by type and half by habitat proportions. In other words, for each option we ascribe a proportion of 0.5 to that option's type category and then multiply its habitat proportions by 0.5 such that the α_{jm} for each option still sum to 1. The final CNMNL model of Table 9 adds membership of mode categories to the type categories and habitat proportions. Again we adjust membership proportions through the assumption that a third is dependent on mode, a third on type and a third on habitat.[5]

With regards to in-sample goodness of fit, the log likelihoods of each of the CNMNL models outperforms the NMNL models for habitat and type similarity groups. One thing to note is that the models tend to fit the estimation data set better when similarity between sites reflects each site's range of habitats rather than just each site's dominant landcover. For example, a site that is mostly woodland but part managed grass is better modelled as being both more similar to sites with woodland and to sites with managed grass than it is as being modelled as being more similar to just other sites with majority woodland landcover.

It is notable from Table 9 that only the CNMNL model with habitat, type and mode similarity groups provides a better fit to the estimation data set than that provided by the NMNL model with just mode categories, an observation that reinforces our initial assessment that accounting for similarity of travel mode options is important in improving in-sample goodness of fit. Indeed, with regards to in-sample fit the best model of those explored is the CNMNL model with habitat, type and mode similarity groups.

<u>Out-of-Sample Goodness of Fit</u>

Of course, the degree to which a model fits the data from which it is estimated is only one measure of its quality. Since the ultimate purpose of the ORVal model is to predict recreational choices perhaps a more important measure of quality is the degree to which the model is able to predict the recreational choices of observations outside the dataset. Indeed, when comparing the performance of different models applied to the same data, out-of-sample validity is considered as important a criterion by which to judge quality as model fit.

Several methods of out-of-sample testing are available. The procedure we adopt here is to draw a second sample of 5,012 observations from those observations in the MENE not selected for the estimation data set. Again we adopt a choice-based sampling strategy to ensure that our testing dataset covers the spectrum of recreational decisions and subsequently reweight our fit statistics to reflect the population. Unlike the estimation dataset, our out-of-sample predictions are not estimated using an importance-sampled choice set. Rather for each observation we include the full choice set including travelling to the 138,621 recreational sites by car and travelling to those same sites on foot. Prediction for each observation, therefore, requires execution of 277,242 routing queries, a time-consuming execution made practical only through application of the contraction hierarchy routing software RoutingKit. As per our in-sample testing, the out-of-sample test statistics

[5] We experimented with specification of the cross-nested logit model in which the proportions attributed to the habitat, type and travel mode similarity sets were estimated from the data. We found those models to be unstable, converging on a corner solution in which the entire weight was apportioned to the set containing the fewest similarity groups. In a model with habitat and type sets the model converged on a maximum where all weight was attributed to the three similarity groups defined in the type set and in a model with habitat, type and mode sets the model converged on a maximum with all weight attributed to the two similarity groups in the travel mode set.

we report are log-likelihoods; that is to say, a measure of the log of the probability our model ascribes to observing the choices made by respondents in the out-of-sample data set.

The final four columns of Table 9 record those out-of-sample fit statistics. The first column records out-of-sample model fit to the recreational choices made by respondents in the out-of-sample dataset. The second column focuses just on the participation decision; that is to say, the success of the model in predicting the daily decisions of respondents as to whether to make an outdoor recreation trip. The third column focuses on the conditional site and mode decision; that is to say, the model's ability to predict the choice over mode and site given a respondent has chosen to taken an outdoor recreational trip. The final column reports out-of-sample fit measures for the conditional choice of mode; that is to say, the choice of which mode of travel to use conditional on having elected to make a trip.

Scanning down the out-of-sample fit statistics for recreational choices the most striking result is that the model demonstrating the best predictive capability is the MNL model with fully specified preference function, a model that demonstrated relatively poor in-sample fit compared to NMNL and CNMNL models. What is more, the model showing the very worst out-of-sample predictive capability is the NMNL with travel mode similarity groups, a model that on the basis of in-sample fit had performed particularly well. Indeed, that poor out-of-sample performance carries over to the CNMNL with habitat, type and mode similarity groups, our best-performing model with regards to in-sample fit.

More clarity over why the NMNL and CNMNL models with mode choice as a similarity grouping perform so badly out-of-sample can be found by studying the out-of-sample predictions for Participation and for Mode/Site choice. The models with 'mode' similarity groups do particularly badly in predicting participation (over-predicting the likelihood of taking a trip) but, for the CNMNL model at least, do quite well at predicting the conditional choice of mode and site.

Of the NMNL models, the model with habitat similarity groups outperforms the other specifications on all out-of-sample measures of goodness of fit. IN terms of out-of-sample fit the CNMNL specification with just habitat similarity groupings performs very similarly to the NMNL model across all measures. Moreover, that relatively simple CNMNL specification is better than the more complex CNMNL specification across most out-of-sample fit measures.

Going forward, therefore, our discussion of parameter estimates will focus on the full MNL model, the NMNL model with habitat similarity groups and the CNMNL model with habitat similarity groups since these specifications appear to provide the best overall performance in predicting out-of-sample recreational choices. Of these three models, CNMNL model gives the best in-sample fit and the MNL the worst with the NMNL taking the middle spot. In contrast, the MNL model gives the best out-of-sample fit and the CNMNL model the worst, with the NMNL model again taking the middle spot. Underpinning the differences in out-of-sample fit is the fact that the MNL model does a better job than the other two models at predicting participation. At the same time, conditional on taking a trip we find that the NMNL and CNMNL do a better job at predicting which site will be visited and by which mode of transport.

Accordingly, no one model can be said to dominate in terms of fit statistics and we turn to an investigation of the parameter estimates to further inform our selection of model.

6.2 Parameter Estimates

We report the parameter estimates in a series of groups. Note that, since the key function of the estimated model is to predict how recreational activity and the welfare it generates might change as the quality and availability of greenspaces is varied, the model was estimated imposing constraints on the signs that could be taken by coefficient estimates. For example, we suppose that adding more expanse to a site of any particular natural land cover cannot decrease the utility offered by that site. Accordingly, we constrain the parameter estimated on areas of habitat cover to be non-negative. Likewise we suppose that endowing a greenspace with some designation should not reduce the benefits it affords visitors and hence we again constrain designation parameters to the positive line.

Pariticpation – Weather

Table 10 is the first of a series of tables reporting parameter estimates for the indirect utility function describing choice of the outside option. Accordingly, positive parameters should be interpreted as indicating that that variable increases the likelihood of choosing the outside option while negative parameters indicate that that variable increases the likelihood of choosing to take a recreational trip. Table 10 focuses on covariates describing the weather at a respondent's home location on each of the seven days for which their recreation decisions are recorded in the MENE data (each day is the unit of choice occasions in our analysis).

Table 10: Parameter estimated for Participation Choices - Weather

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Rain (mm per day)	0.0038*** (3.038)	0.0035*** (2.843)	0.0036*** (2.873)
Rain Squared	-0.0126 (-1.562)	-0.0113 (-1.41)	-0.0115 (-1.445)
Temperature (mean daytime)	-0.0163** (-2.407)	-0.0146** (-2.183)	-0.0142** (-2.112)
Temperature Squared	0.0701 (0.339)	0.0052 (0.025)	-0.0124 (-0.061)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

The parameters for each of the three models paint a broadly similar picture, respondents tend to choose the outside option significantly more often the greater the quantity of rain. In contrast, participation in outdoor recreation increases with temperature. The squared terms on the two weather measurements prove not to be significant in any of the three models.

Participation – When?

Table 11 reports parameter estimates on covariates capturing temporal influences on participation. The first parameter, labelled 'constant' captures the baseline likelihood of choosing the outside option. Notice these parameters are large in absolute magnitude suggesting that, holding everything else constant, respondents tend to choose the outside option significantly more frequently than they choose to take a recreational trip to an outdoor location. Notice that across the three models the magnitude of the constant falls as we move to models that allow for increasing levels of similarity between options. The mathematical underpinning of that observation is that allowing for similarity acts so as to effectively reduce the apparent size of the choice set – in the extreme a group with a very large similarity parameter is regarded by respondents as a single option (each of the options in that group are perfect substitutes for one another). Notice from (6) that the probability of choosing the outside option takes the form of a ratio; the ratio relating the exponentiated utility of the outside option to the exponentiated utilities summed for all sites in the choice set. Accordingly, the greater the level of similarity in options the smaller the sum in the denominator of that ratio and hence, the smaller the constant required to capture preferences for the outside option,

For the remaining covariates the parameters for each model paint a broadly similar picture.

- Participation is significantly higher on Bank Holidays.
- Once accounting for weather and other factors, participation tends to be highest in February, March and June and lowest in September, November and December.
- Likewise, participation is highest on Saturdays and Sundays, and lowest on Mondays.

Table 11: Parameter estimates for Participation Choices - When?

Parameter	**MNL**	**NMNL (Habitat)**	**CNMNL (Habitat)**
Constant	9.0778*** (30.286)	8.761*** (19.725)	8.0589*** (28.751)
Bank Holiday	-0.2564*** (-6.876)	-0.2588*** (-7.025)	-0.2589*** (-7.016)
2009	0	0	0
2010	0.1902*** (4.841)	0.1942*** (5.008)	0.1931*** (4.972)
2011	0.1267*** (3.14)	0.1249*** (3.184)	0.1289*** (3.275)
2012	0.0765* (1.881)	0.0783** (1.983)	0.0818** (2.063)
2013	0.0552 (1.363)	0.051 (1.299)	0.0546 (1.386)
2014	0.0307 (0.782)	0.0325 (0.854)	0.0363 (0.95)
2015	0.0093 (0.235)	0.0151 (0.392)	0.021 (0.544)
2016	0.1629** (2.103)	0.1733** (2.273)	0.1818** (2.373)
Jan	-0.1364** (-2.51)	-0.1269** (-2.36)	-0.1287** (-2.389)
Feb	-0.1946*** (-3.51)	-0.1859*** (-3.392)	-0.1869*** (-3.406)
Mar	-0.1685*** (-3.146)	-0.1657*** (-3.13)	-0.1682*** (-3.173)

Apr	-0.1622*** (-2.971)	-0.1569*** (-2.91)	-0.1579*** (-2.923)
May	-0.0972* (-1.714)	-0.0947* (-1.689)	-0.0954* (-1.699)
Jun	-0.1574*** (-2.647)	-0.1473** (-2.505)	-0.1487** (-2.525)
Jul	-0.0979 (-1.574)	-0.0851 (-1.384)	-0.0871 (-1.415)
Aug	-0.0979 (-1.594)	-0.09 (-1.483)	-0.0899 (-1.479)
Sep	-0.0481 (-0.794)	-0.0374 (-0.624)	-0.044 (-0.733)
Oct	-0.1224** (-2.127)	-0.1182** (-2.081)	-0.1201** (-2.108)
Nov	-0.0494 (-0.893)	-0.0431 (-0.789)	-0.0439 (-0.801)
Dec	0	0	0
Mon	0.5635*** (32.925)	0.5595*** (32.95)	0.5603*** (32.95)
Tue	0.6372*** (37.526)	0.6321*** (37.522)	0.6331*** (37.528)
Wed	0.5992*** (35.589)	0.5944*** (35.588)	0.5953*** (35.591)
Thu	0.5651*** (33.846)	0.5608*** (33.856)	0.5616*** (33.857)
Fri	0.5558*** (33.421)	0.5517*** (33.437)	0.5524*** (33.438)
Sat	0.23*** (14.951)	0.2283*** (14.965)	0.2286*** (14.963)
Sun	0	0	0

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Figure 5 illustrates the parameter estimates for the days of the week variables transformed so as to indicate preferences to take an outdoor recreation trip. To be clear, Figure 4 plots the negative of the parameters rescaled to make Tuesday (the least preferred day for outdoor recreation) the base case. Accordingly, the bars in the Figure provide an indication of the relative magnitude of preferences for taking a trip on each day of the week. As might be expected, people prefer to take trips on the weekend than mid-week.

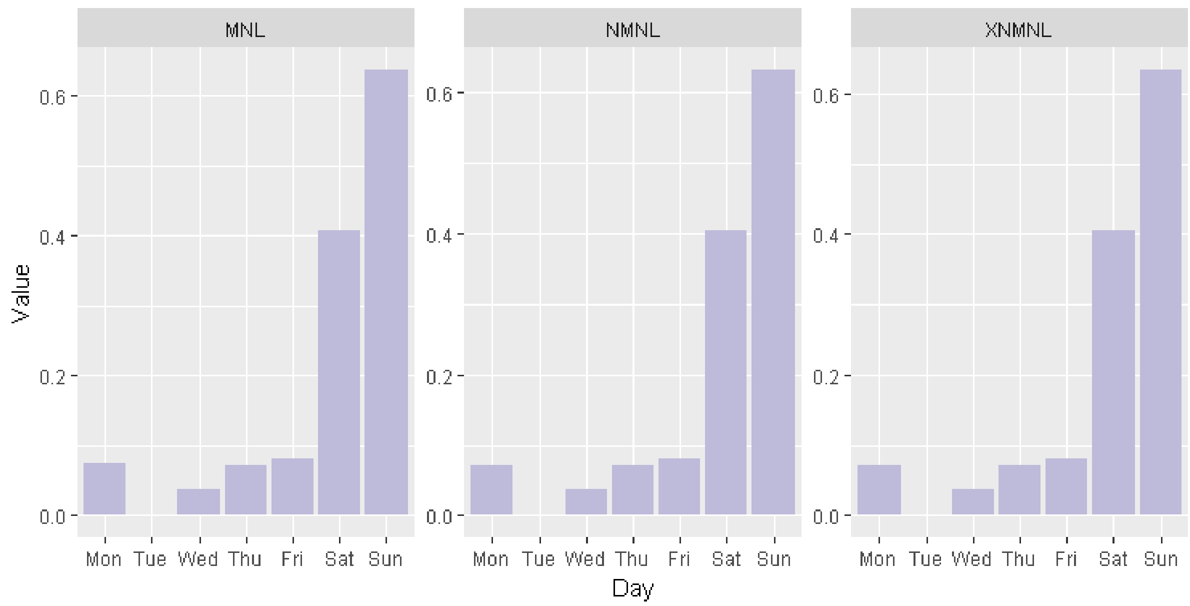

Figure 4: Relative magnitude of preferences to take an outdoor trip on different days of the week (base case Tuesday)

Figure 5 provides a similar illustration for the models' parameters regarding month of the year. Again the plots transform the coefficients from preferences for the outside option to preferences for taking a recreation trip, with the month of December acting as the base case. In this case the pattern of preferences is not as obviously intuitive as those for days of the week; with preferences apparently highest for trips in February, March, April and June. Of course, the key thing to bear in mind is that these are preferences once other factors such as rainfall, temperature and bank holidays have been accounted for.

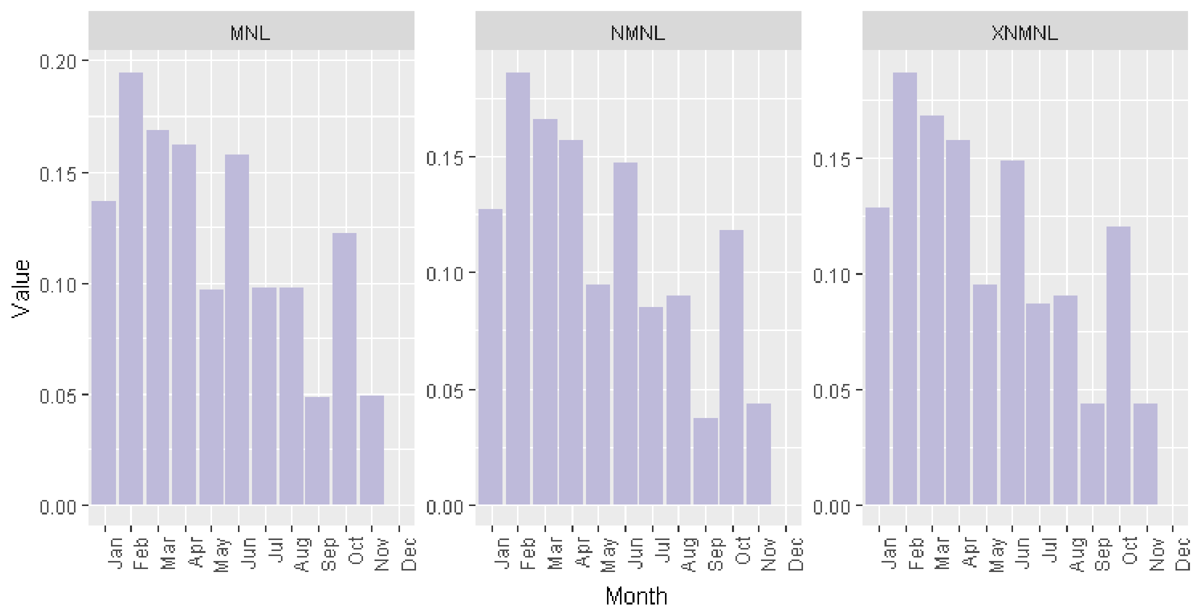

Figure 5: Relative magnitude of preferences to take an outdoor trip during different months of the year (base case February)

Participation – Where?

Table 12 reports parameter estimates on covariates capturing how a respondent's home location impacts on participation. Remember these are independent effects. That is to say they reflect the effect of these covariates once other differences such as patterns of weather and the accessibility of greenspaces have been accounted for. We observe that for all three models, participation in outdoor recreation is significantly lower amongst residents of urban areas, perhaps reflecting the greater availability of alternative forms of recreational activity available in urban areas compared to rural areas. We also observe significant differences in outdoor recreation participation across the English regions. Participation is highest in the North East and South West, and significantly lower in London.

Table 12: Parameter estimates for Participation Choices - Where?

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Urban	0.5478*** (19.842)	0.5006*** (18.526)	0.5063*** (18.693)
North East	0	0	0
North West	0.4601*** (8.345)	0.2694*** (4.985)	0.3098*** (5.696)
Yorkshire and The Humber	0.3804*** (6.713)	0.2049*** (3.685)	0.2098*** (3.746)
East Midlands	0.4035*** (6.871)	0.167*** (2.887)	0.151*** (2.612)
West Midlands	0.4144*** (7.222)	0.1639*** (2.881)	0.1627*** (2.87)
East of England	0.2968*** (5.325)	0.0718 (1.315)	0.0572 (1.05)
London	0.947*** (16.733)	0.5669*** (10.015)	0.5865*** (10.462)
South East	0.3725*** (7.028)	0.1438*** (2.749)	0.139*** (2.675)
South West	-0.0172 (-0.318)	-0.1341** (-2.532)	-0.1496*** (-2.83)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Participation – Who?Table 13 lists parameter estimates on covariates describing how the characteristics of individuals impact on participation in outdoor recreation.

One very clear result is the highly significant impact of dog ownership on outdoor recreation participation; people with dogs take more trips to outdoor greenspace than those without. In contrast, we do not observe a significant increase in participation amongst individuals with children.

Table 13: Parameter estimates for Participation Choices - Who?

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat & Type)
Dog	-1.4695*** (-65.315)	-1.454*** (-65.536)	-1.4562*** (-65.47)
Children	-0.0157 (-0.611)	-0.0161 (-0.636)	-0.0169 (-0.666)
Segment A	-0.6779*** (-11.471)	-0.7098*** (-12.184)	-0.7107*** (-12.171)
Segment B	-0.5687*** (-12.67)	-0.5991*** (-13.572)	-0.6013*** (-13.595)
Segment C1	-0.3578*** (-8.143)	-0.3809*** (-8.805)	-0.384*** (-8.859)
Segment C2	-0.1976*** (-4.308)	-0.2132*** (-4.719)	-0.2152*** (-4.754)
Segment D	-0.06 (-1.244)	-0.0698 (-1.465)	-0.0695 (-1.457)
Segment E	0	0	0
White	0	0	0
Mixed	0.1402 (1.635)	0.1263 (1.483)	0.1216 (1.43)
Black	0.5039*** (8.678)	0.4876*** (8.426)	0.4854*** (8.374)
Asian	0.6464*** (13.738)	0.6274*** (13.361)	0.6282*** (13.356)
Other	0.2718** (2.187)	0.2618** (2.126)	0.2642** (2.14)
Not work	0	0	0
Part time work	-0.0177 (-0.5)	-0.0253 (-0.725)	-0.0255 (-0.728)
Full time work	0.2224*** (7.814)	0.2070*** (7.362)	0.2093*** (7.426)
Female 16-23	0	0	0
Male 16-24	-0.1640*** (-2.991)	-0.1589*** (-2.934)	-0.1596*** (-2.941)
Female 25-34	-0.4328*** (-8.23)	-0.4411*** (-8.483)	-0.4442*** (-8.523)
Male 25-34	-0.3777*** (-6.938)	-0.385*** (-7.152)	-0.3877*** (-7.186)

Female 35-44	-0.414*** (-7.744)	-0.4303*** (-8.139)	-0.4272*** (-8.063)
Male 35-44	-0.3873*** (-6.939)	-0.3982*** (-7.212)	-0.4004*** (-7.237)
Female 45-54	-0.3511*** (-6.585)	-0.3634*** (-6.889)	-0.3637*** (-6.878)
Male 45-54	-0.368*** (-6.656)	-0.3821*** (-6.991)	-0.3822*** (-6.98)
Female 55-64	-0.4083*** (-7.416)	-0.4191*** (-7.695)	-0.4221*** (-7.739)
Male 55-64	-0.4012*** (-7.176)	-0.4221*** (-7.645)	-0.4225*** (-7.64)
Female 65+	-0.1535*** (-2.822)	-0.163*** (-3.028)	-0.1618*** (-2.999)
Male 65+	-0.3646*** (-6.712)	-0.3801*** (-7.069)	-0.3813*** (-7.082)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

As illustrated in Figure 6, across all three models the coefficients estimated on socioeconomic segment show a clear pattern of increasing participation moving from segment E through to segment A. Note that these estimates are the independent effect of socioeconomic segment once we have controlled for a variety of confounding factors such as accessibility to recreation areas and access to a motor vehicle. Accordingly, the models provide evidence to support the idea that outdoor recreation has the properties of what economists call a 'luxury good'.

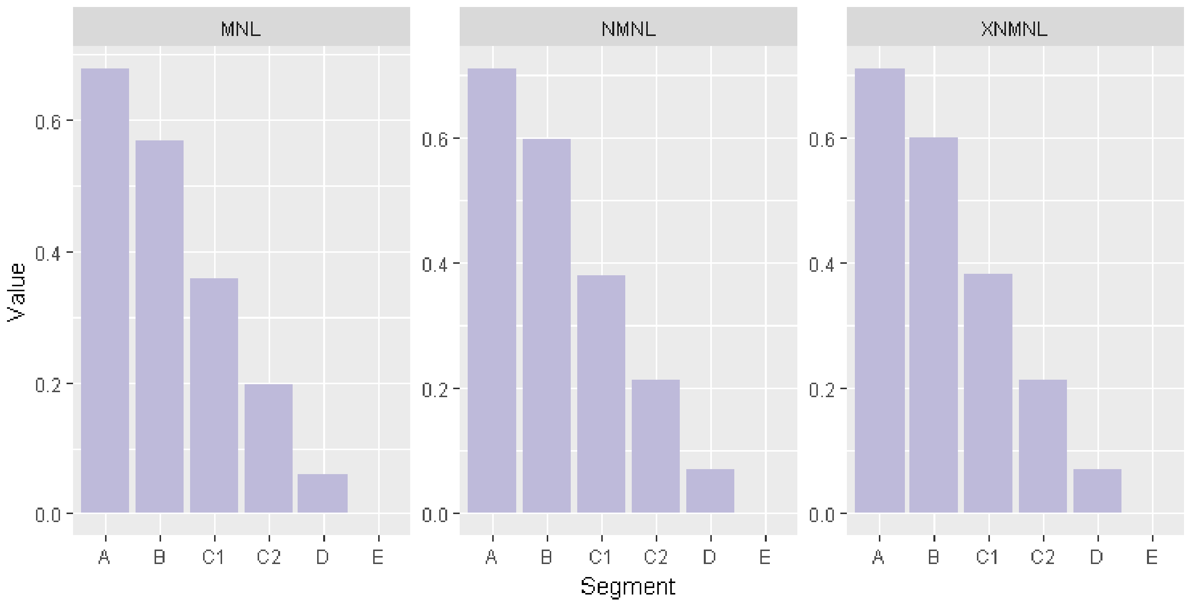

Figure 6: Relative magnitude of preferences to take a trip across socioeconomic groups (base case segment E)

We also observe a strong independent effect from ethnicity; that is to say, having controlled for many other socioeconomic characteristics and the accessibility of greenspace, we find that white respondents are more likely to engage in outdoor recreation than respondents of mixed race who are more likely to participate than black respondents with Asian respondents the least likely to make use of outdoor greenspace.

Significant differences in preferences for participation are also observed across working status with full time workers least likely to participate. Likewise we observe interesting patterns of difference across gender and age groups. Figure 7 illustrates those patterns with the coefficients transformed from preferences for the outside option to preferences for taking a recreation trip. In this case the base case is females aged 16 to 24. Indeed it is that age group and particularly amongst females that we observe the lowest preferences for participating in outdoor recreation. Between the ages of 25 and 64 preferences for participation are reasonably similar and also little different across men and women. Interestingly, amongst the over 65s while participation preferences remain little changed amongst men, we observe a sharp decline in participation in females.

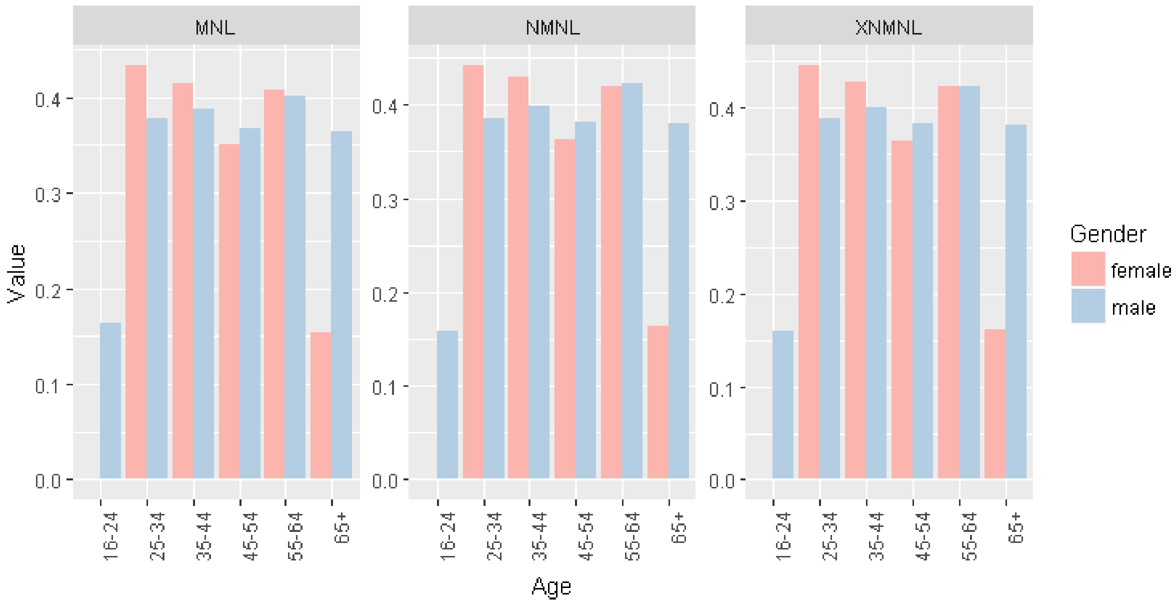

Figure 7: Relative magnitude of preferences to take a trip across age-gender groups (base case females 16-24)

<u>Site and Mode Choice – Travel</u>

Table 14 is the first table listing covariates included in the model to describe choices over mode of travel and site. The parameters shown in Table 14 focus on covariates relating to travel to sites.

Observe that the dummy variable identifying walking options (in contrast to driving options) are positive and significant. Holding everything else constant (including the costs of travel) people are more likely to choose to walk to a greenspace for the purposes of recreation than they are to drive. The model specification also interacts car options with a dummy variable indicating respondents who do not have access to their own private vehicle. Not surprisingly we observe the coefficient on that interacted variable to be negative and highly significant. People without their own vehicle are very much less likely to travel to a recreation site by vehicle.

The parameters on the travel cost for car journeys and walking journeys are, as expected, negative and highly significant; people tend to choose to visit sites that entail less travel. Notice that the parameter on walking travel costs is around 3 times smaller than that on car travel. We conclude that our assumptions regarding the cost of travel time in walking placed too high a cost on walking time.

Finally, in line with the recommendations of the project steering group we introduce separate parameters to account for preferences for travel to allotments. The reasoning behind that specification is that, unlike the other recreation sites, allotments are essentially private goods rationed through an allocation process that restricts access to allotments to certain individuals. As a result of that fundamental difference, our specification is such that a full set of parameters are used to capture preferences for trips to allotments (and in the nested logit model specifications a separate similarity group) that are separate from the parameters used to capture preferences for access to other recreation sites. As it turns out the parameters on the costs of travel to allotments suggests that trips to allotments are much more frequently taken on foot and that that travel is seen as less costly than travel to other types of outdoor recreation site. One might speculate that those parameters suggest that that owners of allotments will tend to choose an allotment within walking distance of their home and that there decision to visit that location is less sensitive to distance than choices between other types of recreation site.

Table 14: Parameter estimates for Site & Mode Choices – Travel

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Car	0	0	0
Walk	3.4002*** (28.814)	2.5194*** (25.729)	2.5769*** (27.015)
Car * No Car	-0.7022*** (-13.727)	-0.7128*** (-15.297)	-0.7188*** (-15.352)
Travel Cost:			
Car	-0.1948*** (-30.214)	-0.1515*** (-28.557)	-0.1608*** (-29.822)
Walk	-0.0611*** (-9.634)	-0.0461*** (-9.361)	-0.0504*** (-9.656)
Travel Cost Allotment:			
Car	-0.5589 (-1.337)	-0.3781 (-1.598)	-0.3707* (-1.653)
Walk	-0.0152** (-2.132)	-0.0137*** (-3.085)	-0.0133*** (-3.555)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Site and Mode Choice – Site Type

Table 15 reports parameter estimates on covariates that differentiate the site options by broad type categories. The first five rows are a dummy variable set with the parks category as the base case and paths, allotments, cemeteries and beaches as alternative site types. The general pattern to the coefficients is that paths and allotments and cemeteries are less preferred to parks while beaches are more preferred. Since these variables pick out dimensions of site characteristics that are also important in defining the similarity groups, it is perhaps not surprising that we now begin to see some differences across the models. Most notable here are the lack of a significant negative coefficient for cemeteries in the NMNL model, and the lack of significant positive coefficient for beaches in the CNMNL model. The importance of these differences in prediction is difficult to ascertain without specific testing.

The specification also includes a variable that records the proportion of a site that is within an urban area, that being the proportion of a parks border that is adjacent to an urban area and the length of a path in an urban area. The MNL returns an insignificant parameter on this variable though those on the NMNL and CNMNL suggest a positive influence with that in the NMNL model being significant at the 95% confidence level.

The final two covariates in Table 15 report parameters on the two commonality factors. Observe that for all three models the commonality factor on paths is negative and highly significant. Recall that the commonality factor records the proportion of the path network accessed from some particular path entry point that is also accessed by other entry points included as path sites in the data. The negative sign on the path commonality factor, therefore, accords with our expectations that a set of path access points serving the same network are considered as close substitutes, thereby, reducing the independent value of each. The commonality factor for parks in the MNL and CNMNL model is also significant but with the opposite (positive) sign. Again that conforms with prior expectations that bordering greenspaces offer complementarities that increase visitation.

Overall the Site Type variables display some differences across the three models, though no one model stands out as resulting in coefficients that better accord with prior expectations. The CNMNL model performs well though has an unexpectedly small coefficient on beaches. The NMNL Model is very similar but returns a positive (though insignificant) coefficient on cemeteries.

Table 15: Parameter estimates for Site & Mode Choices – Site Type

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Parks	0	0	0
Paths	-0.3794 (-0.863)	-1.2201* (-1.824)	-0.6376** (-2.224)
Allotments	-1.2128 (-1.437)	0.1818 (0.238)	-0.4831 (-0.735)
Cemeteries & Graveyards	-0.7103*** (-2.71)	0.0251 (0.051)	-0.7714*** (-2.654)

Beaches	0.8530** (2.175)	0.8036** (2.198)	0.4877 (1.412)
Urban (%)	0.0059 (0.122)	0.0879** (2.402)	0.0703* (1.867)
Commonality Factor – path	-0.6318*** (-12.432)	-0.4278*** (-11.41)	-0.4995*** (-13.129)
Commonality Park - park	0.0358** (2.546)	0.0143 (1.473)	0.0199* (1.959)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Site and Mode Choice – Habitats

One of the key functionalities of the ORVal tool is to predict the values and visits that might be generated by a new recreation site established in a particular location with particular landcovers. Accordingly, a crucial feature of the model underpinning those predictions is that it demonstrates plausible sensitivity to the extent of different habitats.

Our data allow for a reasonably detailed description of habitats each site being described by the extent of moors & heathland, natural grassland, grassland managed for recreation (such as might be found in a municipal park), wetlands including fens & marshes, and woods whose extent is divided into standard woods and wood pasture (open woodland with dispersed trees set in a grass land setting). As discussed previously, our specification of the choice function allows preferences for each site to be determined by a measure of the extent of land under a particular habitat at a site. In particular, for each site the extent of habitat type r is captured through the variable;

$$area\ variable_r = \frac{habitat\ area_r}{\sum_q habitat\ area_q} \times \ln\left(\sum_q habitat\ area_q\right) \qquad (\forall r)$$

that is to say, we examine the effect of the habitat composition of each site by including a set of variables that indicate the share of the log of the total site area under each habitat. Notice that in accordance with our prior expectations the variable is increasing with the addition of an extra unit of a habitat to a site but the larger the total site area the smaller the increment realised from the addition of that extra unit.

The preference function specification also includes a set of dummy variables that identify the dominant habitat type in each site; that is to say, the habitat with the greatest share of a site's total area. The moors & heathland habitat is taken as the baseline category in that set of dummy variables.

Since it may be possible that preference for habitats differs across different types of recreation site. We estimate separate parameters for park- and path-type recreation sites. Table 16 reports parameter estimates on covariates describing the habitat landcovers found at park-type sites.

Table 16: Parameter estimates for Site & Mode Choices – Habitats in Parks

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat & Type)
Moors & Heath:			
constant	0	0	0
area	0.4447*** (6.386)	0.293*** (5.087)	0.1101** (1.994)
Natural Grass:			
constant	-0.6530** (-2.311)	-0.0238 (-0.047)	0.0688 (0.285)
area	0.4198*** (7.101)	0.3272*** (7.417)	0.3275*** (7.328)
Recreational Grass:			
constant	-0.6878*** (-2.722)	1.5461*** (3.404)	0.1621 (0.718)
area - general grass	0.3484*** (19.652)	0.2159*** (16.324)	0.2228*** (15.906)
area - sports pitches	0.1912*** (3.652)	0.1074*** (2.943)	0.1472*** (3.577)
area - formal gardens	0.5910*** (9.314)	0.4077*** (8.58)	0.4219*** (8.572)
Wetlands:			
constant	1.0082** (2.213)	0.5038 (0.828)	0.3876 (1.143)
area	0.2155 (1.193)	0.1772 (1.282)	0 (.)
Woods:			
constant	-0.6693*** (-2.656)	1.0112** (2.225)	-0.0081 (-0.036)
coniferous	0	0	0
broadleaf	0.0989* (1.96)	0.0846** (2.307)	0.0976** (2.427)
felled or young	-0.8044*** (-3.297)	-0.5451*** (-3.178)	-0.5271*** (-2.921)
area - woodland	0.4255*** (20.575)	0.3105*** (19.066)	0.3263*** (19.279)
area – wood pasture	0.6122*** (24.296)	0.4417*** (21.432)	0.4792*** (25.133)
Allotment:			

constant	-1.2128 (-1.437)	0.1818 (0.238)	-0.4831 (-0.735)
area	0.3406** (1.985)	0.1895* (1.711)	0.1927* (1.721)
Cemetery:			
constant	-0.7103*** (-2.71)	0.0251 (0.051)	-0.7714*** (-2.654)
area	0 (.)	0 (.)	0 (.)
Park Habitat Diversity	-0.4506*** (-4.413)	-0.3179*** (-4.291)	-0.1113 (-1.566)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

The first thing to note from Table 16 is that for the MNL and NMNL the continuous covariates on habitat extent are mostly positive and significantly different from zero. As we might hope these models suggest that the marginal value of habitat is positive; in other words, extra habitat at a site has a positive impact on the utility derived from visiting that site. The only exception is the area of cemeteries and graveyards where all three models record the corner solution value of zero that is imposed by our use of constrained optimisation. With the CNMNL model, the parameter on wetland extent also results in a zero coefficient.

The ordering of the coefficients for the extent of the various habitats is reasonably similar across the models. Wood pasture records the highest coefficient suggesting that all else equal the addition of an extra unit of this habitat increases the value derived from a recreational site by the most. The marginal values of heathland and natural grassland are the next highest, with woodland, and recreational grassland returning somewhat smaller values. The lowest marginal values for habitat extent are recorded by allotments, wetlands and sport pitches.

Of course, the value the model specification attributes to a park as a result of the different habitats it offers to recreationists is also dependent on the habitat constants. To help interpret the combined effect of the constants and area coefficients presented in Table 16, Figure 8 plots out the utility value delivered by parks with increasing extents of the different landcovers. To be clear, each of the lines plotted in Figure 8 can be thought of as representing the utility values delivered by a single-habitat park of increasing area.

While differences in scaling of the different models makes it difficult to compare the absolute values of the preferences plotted out in Figure 8, it is apparent that the ordering over the values associated with different habitats differs across the three models. Perhaps the NMNL model returns the most intuitively pleasing ordering, with recreational grassland, woods and wood pasture offering relatively higher levels of value compared to moors & heath, wetlands and natural grassland.

Notice also from Table 16 that our models examine whether the types of trees in woodlands influenced preferences. For parks we find that broadleaf woods are significantly preferred to coniferous woodland, while land that has been newly felled or planted with young trees returns a significant negative coefficient.

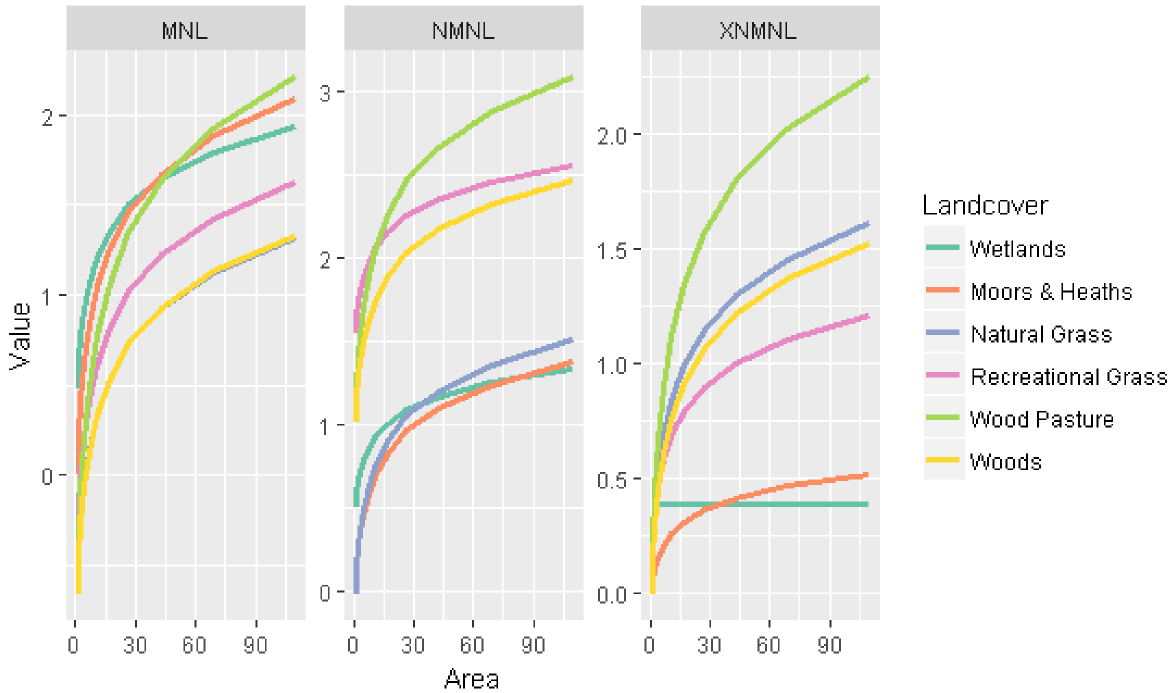

Figure 8: Value of Area of Different Landcover Types in Parks

Table 17 and Figure 9 repeat the analysis of habitat for path-type sites. Once again, the MNL and NMNL models return a set of coefficients that conform with expectations that value should increase with increasing extents of different habitats. For moorland and wetlands the CNMNL model, however, records parameter estimates that are not increasing with increasing area. Indeed, these parameters converge on the corner solution of zero enforced by our use of constrained optimisation.

Table 17: Parameter estimates for Site & Mode Choices – Habitats in Paths

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Moors & Heath:			
constant	0	0	0
area	0.1903 (1.413)	0.1682 (1.451)	0 (.)
Natural Grass:			
constant	-0.3663 (-0.933)	0.9693 (0.975)	0.3838 (1.146)
area	0.5267*** (5.008)	0.3659*** (4.708)	0.4424*** (5.024)
Recreational Grass:			
constant	-0.4044 (-1.136)	2.5929*** (2.672)	0.3458 (1.124)

area	0.3631 (0.879)	0.3135 (1.346)	0.0564 (0.253)
Wetlands:			
constant	0.4785 (0.452)	0.5787 (0.417)	-0.0228 (-0.022)
area	0.2604 (1.005)	0.1866 (0.892)	0 (.)
Woods:			
constant	-0.3067 (-0.849)	2.129** (2.184)	0.4279 (1.36)
coniferous	0	0	0
broad Leaf	0.036 (0.438)	0.0707 (1.169)	0.0804 (1.285)
felled or young	0.3528 (1.212)	0.2391 (1.081)	0.2277 (1.028)
area - woodland	0.2831*** (4.133)	0.2328*** (4.592)	0.3247*** (5.828)
area - wood pasture	0.654*** (4.758)	0.5035*** (4.882)	0.6168*** (5.629)
Agriculture:			
constant	-0.4316 (-1.227)	2.2137** (2.31)	0.3811 (1.229)
area	0.3404*** (5.689)	0.2692*** (6.075)	0.4416*** (9.184)
area - HLS	0 (.)	0 (.)	0 (.)
Path Habitat Diversity	-0.0092 (-0.083)	-0.0203 (-0.288)	0.1329** (2.418)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Considering Figure 9, it is again the NMNL model that returns the perhaps the most intuitively plausible ordering of values. In that model, paths through recreational grassland and wood pasture return the highest values. Agricultural land, natural grassland and woodland return the next highest values with wetlands and moors & heathland providing the least value. For paths through woodland, the parameters on tree type and age show no significant effect of these factors on preferences.

Overall, the models demonstrate that habitat extent is a significant determinant of site choice. With regards to sensitivity to habitat extent and the plausibility of the relative size of preferences for different habitats, the NMNL model returns coefficients that are perhaps most in tune with prior expectations.

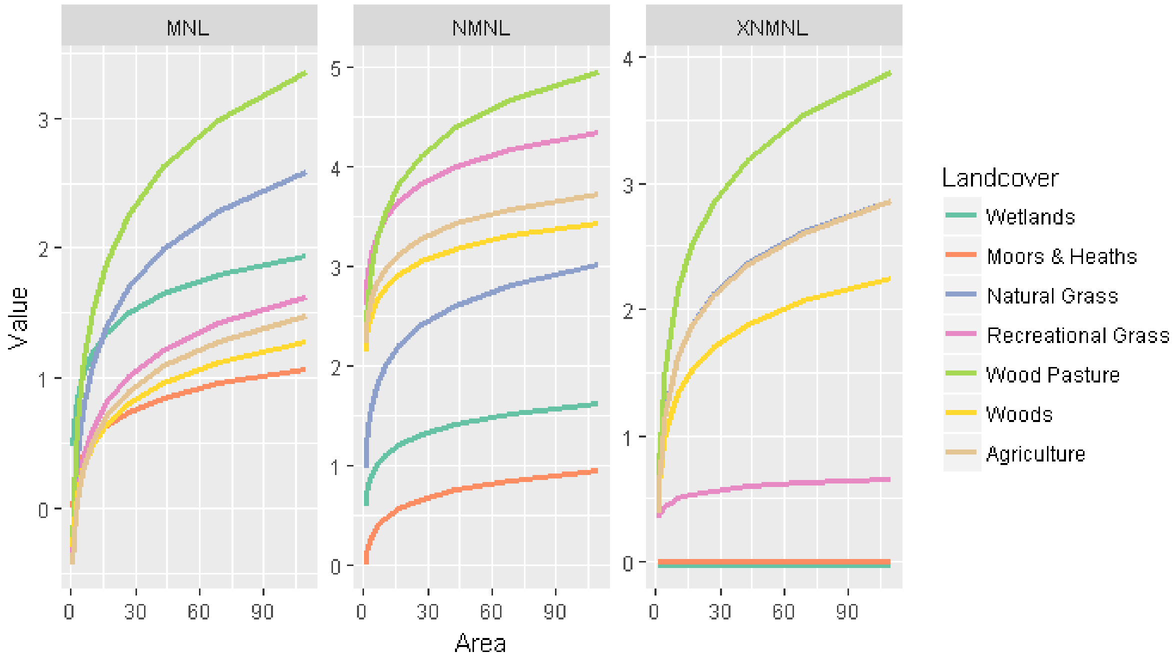

Figure 9: Value of Area of Different Landcover Types for Paths

Site and Mode Choice – Water

In addition, to habitat extent, the models include a detailed description of the extent of different water features to be found at each site. Again we use a specification in which the composition of water features is included through a series of variables capturing the extent of different types of water body. Indeed, for each type of water body we adopt the same specification as used for land habitats; namely the share of the log total area of water features comprised of that water body type.

At the top level we distinguish between freshwater and saltwater features, including dummy variables to distinguish sites where either of those features are the dominant habitat to be found at a site.

Amongst freshwater features we include extent of rivers & canals as well as extent of lakes. For rivers & canals we also include a dummy variable identifying rivers where the WFD ecological status variable is in the 'good' or 'excellent' category. Amongst saltwater feature we include extents of sea, estuaries and saltmarshes to be found at each site. Again we estimate separate parameters for park- and path-type recreation sites.

Table 18 reports parameter estimates describing preferences for the waterscapes of park-type recreation sites. Figure 10 visualises those estimates plotting out the utility value delivered by parks with increasing extents of different water bodies.

For all three models we observe that increasing size of lakes & reservoirs in park-type sites has a strong positive effect on values. Of the other water features, the NMNL model shows the most by way of the expected sensitivity to waterscapes with estuaries, rivers & canals and saltmarshes also returning positive coefficients though these are not statistically significant.

Table 18: Parameter estimates for Site & Mode Choices – Water Features in Parks

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Freshwater			
constant	-0.0588 (-0.224)	0.9523** (2.014)	0.0347 (0.141)
area - lake & reservoir	0.2282*** (6.596)	0.1728*** (6.745)	0.1613*** (6.075)
area - river & canal	0 (.)	0.0152 (0.751)	0.0028 (0.134)
low ecological quality	0	0	0
high ecological quality	0.3714*** (5.113)	0.2445*** (4.51)	0.2329*** (4.178)
Saltwater			
constant	1.2914*** (3.645)	0.9809* (1.878)	0.2337 (0.726)
area - sea	0 (.)	0 (.)	0 (.)
area - estuary	0.168 (1.429)	0.0708 (0.604)	0 (.)
area - saltmarsh	0 (.)	0.0219 (0.146)	0.002 (0.017)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

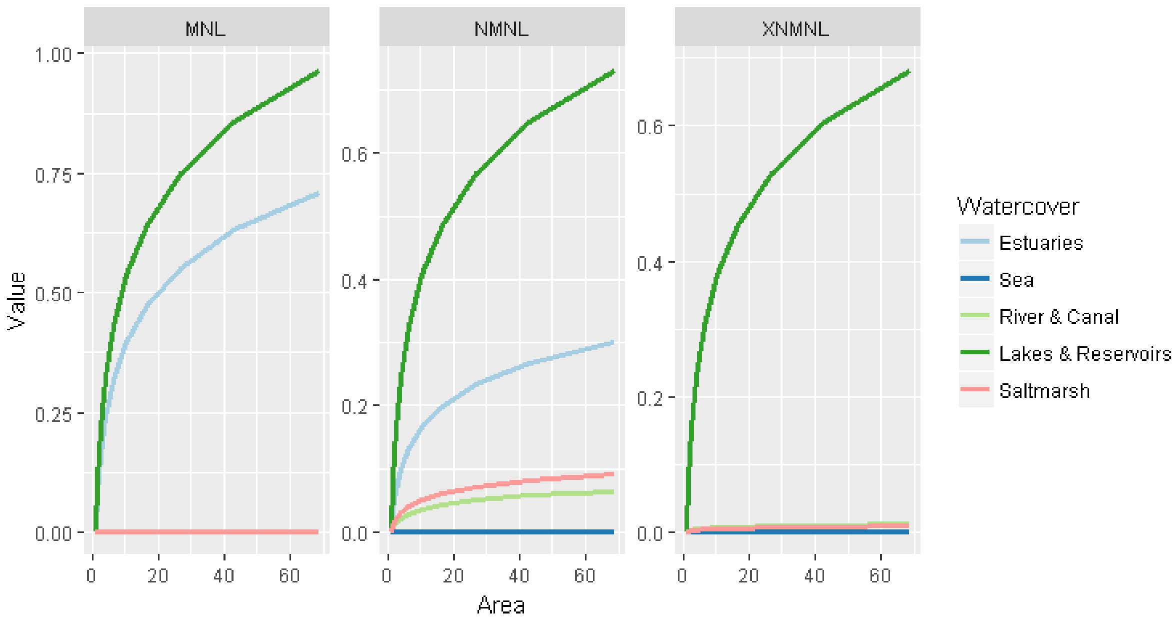

Figure 10: Value of Area of Different Watercover Types for Parks

The same analysis for water features accessed by path-type sites are reported in Table 19 and Figure 11. Notice how, in contrast to park-type features, we see a far greater sensitivity to extent of water features for path-type sites. Across the models we see significant positive coefficients on the extents of Rivers & Canals, Lakes, Sea and Estuaries. One possibility to explain that difference is that access to water features is often best served by a path-type recreation site that follows the border of the water feature; for example, along a river or coastal path. We suspect that waterscapes are more likely to be the key feature of path-type recreation sites than they are of park-type sites.

Table 19: Parameter estimates for Site & Mode Choices – Water Features in Paths

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Freshwater			
constant	0.3488 (0.734)	1.1192 (1.123)	0.3785 (1.003)
area - lake & reservoir	0.1367 (1.602)	0.1161* (1.883)	0.0299 (0.503)
area - river & canal	0.1979*** (5.541)	0.1171*** (4.567)	0.0139 (0.525)
low ecological quality	0	0	0
high ecological quality	0.3714*** (5.113)	0.2445*** (4.51)	0.2329*** (4.178)
Saltwater			
constant	0.0458 (0.125)	1.8833* (1.926)	0.5822* (1.765)
area - sea	0.4585** (2.305)	0.3160** (2.544)	0.1155 (0.979)
area - estuary	0.4513** (2.199)	0.1850 (1.113)	0 (.)
area - saltmarsh	0.2818 (0.62)	0.0412 (0.15)	0 (.)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

Across all three models we find that paths that follow the coast are most highly valued. In the MNL model, the extent of access to estuaries is valued very similarly to access to the coast. In the NMNL model estuary values are somewhat smaller, but in the CNMNL model access to estuaries converges on the corner solution of zero value. Within each of three models we observe that the extent of access to rivers and to lakes & reservoirs are attributed very similar values. In the NMNL those values are both statistically significant at the 90% level of confidence or higher whereas in the CNMNL neither prove to be significant. In all models high river water quality is positively valued.

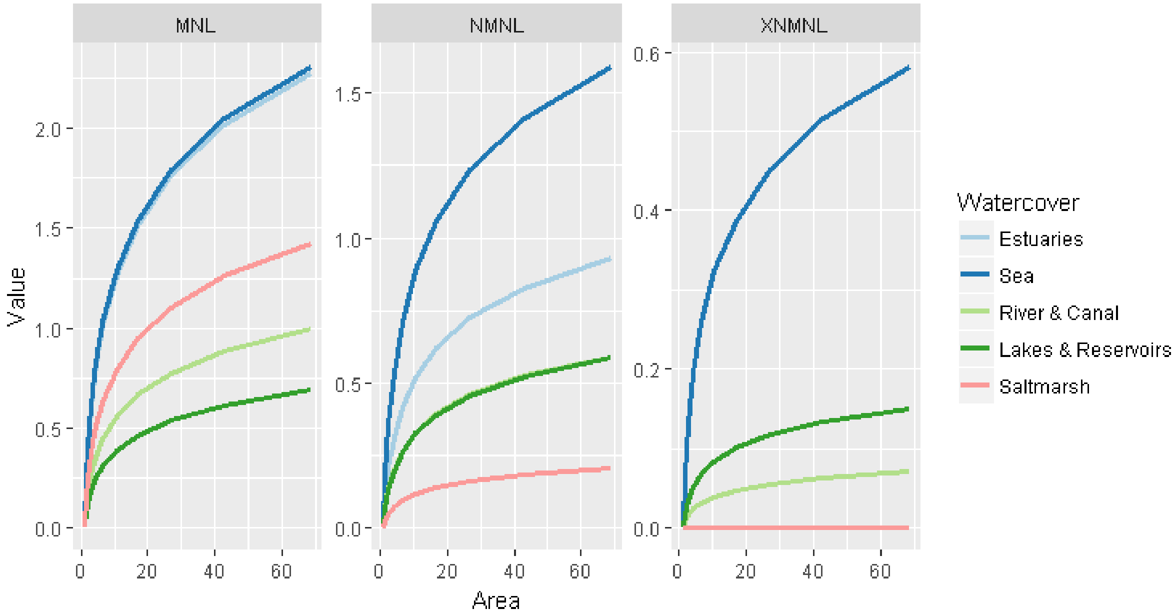

Figure 11: Value of Area of Different Watercover Types for Paths

Overall, it is again the NMNL model which gives an ordering to preferences that is perhaps most plausible; with access to sea seen as the most valuable water feature, estuaries somewhat less valuable, followed by almost identical preferences for rivers & canals and lakes & reservoirs with saltmarsh appearing as the least valuable water feature.

Site and Mode Choice – Beaches

Table 20 reports parameter estimates on covariates describing the qualities of beaches. For all three models beaches with high water quality are preferred to those with low quality or sand, shingle or sand & shingle are valued more greatly than beaches that are predominantly rocky or of solid man-made construction. Perhaps the most noticeable difference between the models Is that both the MNL and NMNL models follow prior expectations signalling a strong positive preference for beach recreation (all else held equal), something that is not returned by the CNMNL model.

Table 20: Parameter estimates for Site & Mode Choices – Beaches

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Beach	0.853** (2.175)	0.8036** (2.198)	0.4877 (1.412)
Low Water Quality	0	0	0
High Water Quality	0.3432*** (2.968)	0.3446*** (3.288)	0.3456*** (3.169)
Rocky or Harbour	0	0	0
Sand	0.4164* (1.657)	0.3827 (1.616)	0.4016* (1.646)

Shingle	0.5317* (1.942)	0.459* (1.776)	0.4715* (1.77)
Sand & Shingle	0.3458 (1.333)	0.3258 (1.328)	0.3429 (1.357)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

<u>Site and Mode Choice – Points of Interest</u>

Table 21 reports parameter estimates on covariates indicating the presence of points of interest at recreation sites. As anticipated for all three models we find that each of the different categories of points of interest are positively regarded. Only the parameters on viewpoints fail to show a statistically significant relationship.

Table 21: Parameter estimates for Site & Mode Choices – Points of Interest

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Archaeology	0.3271*** (4.712)	0.2065*** (3.985)	0.199*** (3.79)
Historic Building	0.2947*** (4.594)	0.2188*** (4.708)	0.2173*** (4.549)
Scenic Features	0.1503* (1.764)	0.139** (2.202)	0.1325** (2.08)
Playground	0.5412*** (16.806)	0.3681*** (15.744)	0.4119*** (17.34)
Viewpoint	0.0226 (0.387)	0.0311 (0.74)	0.0061 (0.142)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

<u>Site and Mode Choice – Designations</u>

Table 22 and **Table 22**Table 23 report parameter estimates on covariates indicating the proportion of a sites extent under different designations for park-type and path-type recreation sites respectively.

Table 22: Parameter estimates for Site & Mode Choices – Park Designations

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
National Park	0.3023*** (5.758)	0.1800*** (3.863)	0.1942*** (4.031)
AONB	0.1412*** (3.561)	0.0946** (2.547)	0.0996*** (2.677)
CROW	0 (.)	0.0079 (0.19)	0.0048 (0.109)

Heritage Coast	0 (.)	0 (.)	0 (.)
Nature Designation	0 (.)	0 (.)	0 (.)
Country Park	0.3732*** (5.667)	0.2790*** (5.686)	0.3079*** (6.226)
Historic	0.1288*** (8.951)	0.0843*** (8.481)	0.0920*** (8.759)
Millennium or Doorstep Green	0.1094** (2.108)	0.0786** (1.999)	0.0792** (1.972)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

For designations in parks, all three models return very similar parameter estimates. Country Park status is accorded the highest value. Parks in National Parks are also attributed high values with sites in AONBs, Millennium Greens and parks with Historic status returning smaller yet still positive and significant coefficients. Across all three models no additional value is associated with parks along Heritage Coasts, those with various nature designations or those designated as CROW.

Table 23: Parameter estimates for Site & Mode Choices – Path Designations

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
National Park	0.5517*** (5.071)	0.3097*** (3.477)	0.2904*** (3.133)
AONB	0 (.)	0 (.)	0 (.)
CROW	0.1865*** (3.078)	0.1457*** (3.996)	0 (.)
Heritage Coast	0.3967 (1.525)	0.0553 (0.29)	0.0655 (0.34)
Nature Designation	0.0148 (0.166)	0.0020 (0.027)	0 (.)
National Trail	0.0557 (0.389)	0.1804* (1.778)	0.1229 (1.203)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. Coefficients significant at the 90% level are highlighted with ***, those at the 95% level with ** and those at 99% at *.

The designation parameters for path-type sites also identify National Parks as settings that deliver higher value. For the NMNL and MNL models, paths passing through areas with CROW designation also show a significant positive effect. Of the remaining designations the only other significant parameter is returned by the NMNL model where paths with National Trails designation are found to be valued positively.

Looking at the designation parameters as a whole it is again the NMNL model which returns patterns of significant parameter estimates that attune most closely with prior expectations.

Similarity Groups

The final set of parameters to be considered are those capturing the level of similarity within similarity groups. Those parameters are listed in Table 24. Of course, no such parameters are estimated for the MNL which does not allow for similarity groupings. Recall also that in the NMNL model sites are assigned exclusively to the similarity group for the dominant habitat at that site whereas for the CNMNL model sites can be members of multiple similarity groups where there membership of a habitat similarity group is determined by the proportion of the site area under that habitat.

For the estimated model to be universally consistent with the assumptions of random utility theory, the similarity parameters should take a value greater than 1. That is true for all similarity parameters in the NMNL model, but for the CNMNL model the parameters on salt water habitat and wetland habitat converge on the value of 1 set as a constraint in the optimisation.

Table 24: Parameter estimates for similarity groups

Parameter	MNL	NMNL (Habitat)	CNMNL (Habitat)
Woodland		1.3717*** (14.383)	1.3848*** (15.874)
Salt Water		1.0362 (0.82)	1 (.)
Fresh Water		1.3310*** (7.083)	1.2575*** (8.187)
Recreational Grass		1.4562*** (20.161)	1.3977*** (23.14)
Agriculture		1.4112*** (13.482)	1.4975*** (19.804)
Natural Grass		1.2559*** (3.905)	1.5703*** (10.899)
Wetland		1.0022 (0.012)	1 (.)
Moors & Heathland		1.0216 (0.149)	1.051 (0.755)
Allotments		1.5023*** (4.568)	1.5082*** (4.486)
Cemeteries		1.2862*** (5.567)	1.3196*** (6.621)

Notes: Statistics report the coefficient estimate with the robust standard error below in brackets. In this case coefficients significance is relative to a value of 1 with those significant at 90% level highlighted with ***, those at the 95% level with ** and those at 99% at *.

7. Conclusions and Model Choice

The ORVal extension project has allowed for a significant programme of model development and testing. That programme of work has included the introduction of a travel mode dimension to choice, the refinement of key covariates (particularly, travel costs), the inclusion of new covariates including weather and measures of habitat quality, and an improved specification of the preference function.

The parameter estimates returned by the new models generally conform to prior expectations regarding impacts of covariates on participation and choice of mode and site. At the same time, there are important differences across specifications of the model allowing for different patterns of similarity between recreation sites, an issue we return to subsequently. Of particular interest given the objectives of the ORVal extension project is the fact that coefficients on measures of woodland quality, bathing water quality and river ecological quality all return significant parameters with expected signs. Moreover, the new ORVal recreation demand models demonstrate markedly improved sensitivity to key determining factors such as land and water cover than was exhibited by the original ORVal model.

A final decision to be made concerns which particular specification of the model to use in developing the online ORVal tool. Two different considerations might inform that choice. First, we would like to use a model that performed well with various statistical measures of goodness-of-fit. Second, we would like a model whose parameters deliver predictions that attune best with prior expectations regarding the impact of variables on recreational activity.

With regards to measures of goodness-of-fit, we find that significant improvements in *within-sample* fit can be achieved through the specification of models with increasingly complex definitions of similarity groupings within the GEV framework. According to within-sample fit, therefore, the CNMNL specifications perform best. In contrast, the MNL model, a specification with no similarity groups, has the best out-of-sample prediction performance. That out-of-sample performance results primarily from the MNL's superior ability to predict participation decisions. Both the NMNL and the CNMNL perform worse in predicting that participation decision but actually perform relatively better on predicting which particular recreation site is chosen for a visit.

While the CNMNL may be preferred on in-sample fit measures and the MNL on out-of-sample fit measures, our review of the coefficient estimates suggests that the NMNL model results in parameters that accord better with prior expectations and are responsive to the range of changes that are central to the ORVal tool functionality. One of those functions, for example, is predictions regarding changes in recreational behaviour resulting from changes in landcover at a particular site. Compared to the CNMNL model in particular, the NMNL model returns the largest array of positive and significant coefficients on the land and water habitat variables implying that model will provide predictions that are responsive to changes in landcover,

One further concern regards the fact that the ORVal tool publishes predictions of the predicted welfare value generated by each site and the predicted number of trips to that site. It is a mathematical reality of that model that as the number of options increases that the ratio of predicted welfare value to predicted visits (that is to say, welfare per visit) will converge on a constant given by the negative inverse of the travel cost parameter; in this case a value per predicted trip of £5.13. The intuition behind that result is that in arriving at its predictions of which site people choose to visit the MNL model imposes the assumption that no site is more similar to some other site than it is to any other site. Since the model's predictions are probabilistic, we find that the model results in a classic equi-marginal outcome in which people are assumed to spread out their probability of visiting across sites in direct proportion to the welfare they will gain from visiting each site. If a site were to offer a ratio of welfare to visitation probability than was greater than other sites, then the optimal response would be to increase the probability of visiting that site until the ratio of welfare to probability of visiting was the same across all sites. Following the same logic, we would expect to see the NMNL returning a welfare per expected visit measure that is the same for all sites in a single similarity group but differs across similarity groups. The welfare to predicted visits measure for the CNMNL model will differ across all parks according to the particular shares of habitat in a park.

To summarise, there is no one model that clearly dominates the others;

- The CNMNL model returns the best in-sample fits and returns estimates of the welfare per visit ratio that differs across sites. At the same time, the CNMNL model does not perform particularly well in terms of out-of-sample fit and returns parameters that show insensitivity to a number of factors that is discordant with prior expectations.
- The MNL logit model performs particularly well in out-of-sample tests particularly as a result of its superior ability to predict the participation decision. The parameter estimates of the model accord reasonable well with prior expectations but result in estimates of the welfare to visits ratio which, counter-intuitively, is identical for every site.
- In terms of goodness-of-fit, the NMNL model does not exceed in any of the measures of within- or out-of-sample prediction. Indeed, it performs largely the same as the CNMNL model according to these measures. At the same time, the NMNL model returns perhaps the largest array of coefficient estimates that are in accordance with prior expectations and at least predicts different welfare per visit values that differ across similarity group.

Since no strong argument arises for any one model, our recommendation is to take forward the NMNL model, recommendation made primarily on that model's superior sensitivity to an array of factors that might be expected to influence choice across sites (in particular, sensitivity to different land and water habitats and to a range of designations).

References

Ben-Akiva, M. and M. Bierlaire (1999). Discrete choice methods and their applications to short term travel decisions. Handbook of transportation science, Springer**:** 5-33.

Bierlaire, M. (2006). "A theoretical analysis of the cross-nested logit model." Annals of Operations Research **144**(1): 287-300.

Cascetta, E., A. Nuzzolo, F. Russo and A. Vitetta (1996). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Pergamon Lyon, France.

Department for Transport (2014). Values of Time and Vehicle Operating Costs: TAG Unit 3.5.6, Department for Transport.

Department for Transport (2015). Provision of Market Research for Value of Travel Time Savings and Reliability, Phase 2 Report, Department for Transport.

Guevara, C. A. and M. E. Ben-Akiva (2013). "Sampling of alternatives in Multivariate Extreme Value (MEV) models." Transportation Research Part B: Methodological **48**: 31-52.

Kling, C. L. and J. A. Herriges (1995). "An Empirical Investigation of the Consistency of Nested Logit Models with Utility Maximization." American Journal of Agricultural Economics **77**(4): 875-884.

Manski, C. F. and S. R. Lerman (1977). "The Estimation of Choice Probabilities from Choice Based Samples." Econometrica **45**(8): 1977-1988.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers of Econometrics. P. Zarembka. New York, Academic Press.

McFadden, D. (1978). Modelling the choice of residential location. Spatial Interaction Theory and Planning Models. A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull. Amsterdam, Amsterdam**:** 75-96.

McFadden, D., W. B. Tye and K. Train (1977). An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model, Institute of Transportation Studies, University of California.

Morey, E. R. (1999). Two RUMs unCLOAKED: Nested-Logit Models of Site Choice and NestedLogit Models of Participation and Site Choice. In Valuing Recreation and the Environment: Revealed Preference Methods in Theory and Practice, Citeseer.

Natural England (2015). Monitor of Engagement with the Natural Environment: Technical Report from year 6 of the survey. **Natural England Joint Report JP015**.

Small, K. A. and H. S. Rosen (1981). "Applied Welfare Economics with Discrete Choice Models." Econometrica **49**(1): 105-130.

Smith, V. K. and Y. Kaoru (1990). "What have we learned since hotelling's letter?" Economics Letters **32**(3): 267-272.

Thiene, M., J. Swait and R. Scarpa (2017). "Choice set formation for outdoor destinations: The role of motivations and preference discrimination in site selection for the management of public expenditures on protected areas." Journal of Environmental Economics and Management **81**: 152-173.

APPENDIX 1: Destination Matching Algorithm

For each observation, i, the scoring procedure progressed through the following steps.

- Potential Sites: To identify greenspaces that were potential matches to the visit destination, all parks, beaches and path networks with a within 2.5km of the location recorded as the MENE destination location were selected from the ORVal Greenspace Map. For beaches and parks the proximity of the potential site was recorded as the straight line distance from the centroid of that site to the MENE destination location. For path networks the proximity was taken as the straight line distance to the nearest location on a path network.
- Location Score: A proximity index was calculated for each site in the list of potential matches (indexed by s) using the following formula:

$$1 - \frac{Proximity_s \text{ (in m)}}{2{,}500} \tag{21}$$

which ascribes an index of 1 to sites exactly on the recorded MENE destination location and declines linearly with distance to 0 for the most distant potential match sites 2.5km from the recorded destination location.
For path networks, sites are defined by access points such that a second round of logic was required. First, we identified all access points to each site network in the list of potential matches. We then ranked those according to how far the access point was from the point we had previously identified as the nearest point on that network to the recorded MENE destination location with rank 0 being the closest, 1 the second closest, 2 the third closest, and so on. Under the assumption that it was more likely that we calculated the proximity index for access point p on path network s as follows;

$$\left(1 - \frac{Proximity_s \text{ (in m)}}{2{,}500}\right) \times 0.95^{rank_{s,p}} \tag{22}$$

Such that the highest ranked path access point on the network was given the highest proximity index and that index declined geometrically with increasing rank.
A *location score* was calculated first by multiplying the proximity index by a positive weighting factor. As with the other weighting factors to be described subsequently this weighting factor was adjusted in a process of calibration that ultimately set its value to 50.
Finally the *location score* for each site was adjusted to reflect information provided by respondents in the MENE questionnaire on the distance they had travelled to get to the site. That information was provided as a range such that if the distance between the respondent's home and a possible match site was less than half the distance of the low end of that range then the proximity score was adjusted by a factor given by;

$$\frac{Distance\ to\ Home}{0.5\ \times Lower\ End\ of\ Distance\ Range} \tag{23}$$

Likewise, if the distance from a respondent's home was greater than 1.5 times the high end of the reported distance range then the proximity score was adjusted by a factor given by;

$$\frac{1.5\ \times Lower\ End\ of\ Distance\ Range}{Distance\ to\ Home} \tag{24}$$

Clearly both adjustment factors lie between 0 and 1 ensuring that possible match sites located at a distance from a respondent's home considerably different from the distance they reported in the MENE questionnaire end up with a lower overall proximity score.
The final *location score* varied on the range between 0 and 50.

- Environs Score: Questions 2 and 5 of the MENE survey provide information that helps identify the environs of the visited site particularly whether it was in a built-up or rural location, and whether that on was coastal or inland. To calculate an *environs score* for each possible match site, we began by defining a built-up indicator variable, *built-up%*, which established the proportion of a park's boundaries or a path's length that was within 100m of a built-up area. Where a respondent answered that they had visited a location in a town, city or seaside resort then we began the calculation of an *environs score* for each possible match site by multiplying a weighting factor (calibrated value: 10) by *built-up%*. Alternatively, if they indicated they had visited a location in the countryside a possible match sites location score was calculated as the weighting factor multiplied by 1 − *built-up%*. A similar calculation was carried out for coastal locations where coastal proximity was turned into a linearly declining index equal to 1 at the coast and falling to a value of 0 5km inland. Again if the visit destination was recorded as coastal then a weighting factor (calibrated value: 10) was multiplied by the coastal proximity index otherwise it was multiplied by one minus that amount. The *environs score* was incremented by that value reflecting the degree to which the coastal environs of the visited site matched that of possible match site.
  Answers to Question 5 of the MENE survey gave further clues as to the environs of the chosen site; for example, a respondent indicating that they had visited "Farmland" was assumed to have visited a site in a rural setting, while those indicating they had visited "A park in a town or city" had clearly chosen a site in a built-up setting. Such confirmatory information was given a weighting factor (calibrated value: 5) and added to the total *environs score*, which as a result could take a maximum value of 25.
- Type Score: Answers to Question 4 and 5 of the MENE survey allowed us to compare the type of recreation site visited by a respondent to the types of the possible match sites. Some explicit responses were given very high weighting factors; for example, if a respondent stated they had visited "an allotment", then all allotments in the list of possible sites were given a *type score* of 50 while all sites that were not allotments were given a *type score* of -10. Where the details of the Question 5 response were less explicit a lower type score was attributed; for example, if a respondent stated they had visited "a playing field or other recreation area" then a *type score* of 8 was given for all possible visit sites classified as 'parks' and a *type score* of 0 to all possible visit sites with a different classification.
- Landcover Score: Similar to the *type score* the landcover score used evidence from Questions 4 and 5 of the MENE survey to establish how closely the sorts of landcovers

present at the possible visit sites matched those present at the site actually visited. As an example, respondents indicating they had participated in fishing, swimming outdoors or watersports must have visited a site bordering water features including rivers, lakes, estuaries and sea. Accordingly, sites with such features amongst the list of possible match sites were given an increased *landcover score*. Similarly, where a respondent indicated they had visited a woodland or forest, then possible match sites with woodland cover were attributed landcover score.

- Total Match Score: To arrive at an overall match score for each possible site, the location score, environs score, type score and landcover score were summed. The site with the highest match score was chosen as the most likely location of that particular focus visit.

The full matching algorithm is transcribed below:

```
CREATE TABLE MENE.NearSites
 (spid   bigint,
  type   varchar,
  supertype varchar,
  prox   float,
  areagrid float,
  urbanpct float,
  coastprox float,
  lc_woods float,
  lc_agrculture float,
  lc_moors_heath float,
  lc_mountain float,
  lc_coastal float,
  lc_wood_pasture float,
  lc_sports_pitches float,
  lc_golf float,
  lc_allotments float,
  lc_seaside float,
  lc_estuary float,
  lc_rivers_canals float,
  lc_lakes_reservoirs float,
  dg_ancient_woodland float,
  dg_sssi float,
  dg_CPark float,
  dg_natura2000 float,
  dg_nnr float,
  dg_lnr float,
  dg_ramsar float,
  poi_playground float,
  disthome float,
  near_score integer,
  loc_score integer,
  type_score integer,
  lc_score integer,
  score  integer);

DO
$$
<<DestinationMatching>>
DECLARE
 destination record;
 visitdata record;
 matches        record;
 proximity integer := 2500;
 counter        integer := 0;
 nprint  integer := 50;
 pcounter      integer := 0;
 numrows        integer;
 numsites      integer;
 matchcnt      integer;
 coastdist     integer := 5000;
 coastthrshld integer := 1000;
 ntoprocess float;
 nprocessed float;
 TEST   boolean := FALSE;
```

```
BEGIN

 TEST = FALSE;

 ntoprocess := (SELECT count(*) FROM mene.visit_match);

 FOR destination IN TABLE mene.visit_match LOOP

   IF destination.geom_dest IS NULL OR destination.geom_home IS NULL THEN
     -- No destination or home location data - MATCHCODE -2
     IF NOT TEST THEN
        EXECUTE 'UPDATE mene.visit_match SET matchcode = -2 WHERE visitid =
'||destination.visitid;
     END IF;
   ELSIF lower(destination.q9) != 'your home' THEN
     -- Not travelled from home - MATCHCODE -3
     IF NOT TEST THEN
        EXECUTE 'UPDATE mene.visit_match SET matchcode = -3 WHERE visitid =
'||destination.visitid;
     END IF;
   ELSIF (lower(destination.q4_06) = 'yes') OR (lower(destination.q4_09) = 'yes') OR
(lower(destination.q4_11) = 'yes') THEN -- Not a location based interaction with greenspace
      -- q4_6: Off-road driving or motorcycling
      -- q4_9: Road cycling
      -- q4_11: Appreciating scenery from your car (e.g. at a viewpoint)
      -- MATCHCODE -4
     IF NOT TEST THEN
        EXECUTE 'UPDATE mene.visit_match SET matchcode = -4 WHERE visitid =
'||destination.visitid;
     END IF;
   ELSIF (lower(destination.q5_05) = 'yes') AND
      (lower(destination.q5_01) = 'no') AND (lower(destination.q5_02) = 'no') AND
(lower(destination.q5_03) = 'no') AND (lower(destination.q5_04) = 'no') AND
      (lower(destination.q5_06) = 'no') AND (lower(destination.q5_07) = 'no') AND
(lower(destination.q5_08) = 'no') AND (lower(destination.q5_09) = 'no') AND
      (lower(destination.q5_10) = 'no') AND (lower(destination.q5_11) = 'no') AND
(lower(destination.q5_12) = 'no') AND (lower(destination.q5_13) = 'no') AND
      (lower(destination.q5_14) = 'no') AND (lower(destination.q5_15) = 'no') AND
      (lower(destination.q4_02) = 'no') AND (lower(destination.q4_03) = 'no') AND
(lower(destination.q4_04) = 'no') AND (lower(destination.q4_05) = 'no') AND
      (lower(destination.q4_06) = 'no') AND (lower(destination.q4_07) = 'no') AND
(lower(destination.q4_08) = 'no') AND (lower(destination.q4_10) = 'no') AND
      (lower(destination.q4_12) = 'no') AND (lower(destination.q4_13) = 'no') AND
(lower(destination.q4_15) = 'no') AND (lower(destination.q4_16) = 'no') AND
      (lower(destination.q4_17) = 'no') AND (lower(destination.q4_18) = 'no') AND
(lower(destination.q4_19) = 'no') THEN
     -- Then just visited a village - MATCHCODE -5
     IF NOT TEST THEN
        EXECUTE 'UPDATE mene.visit_match SET matchcode = -5 WHERE visitid =
'||destination.visitid;
     END IF;
   ELSE

    -- (1) FIND NEAR SITES
    -- -------------------
    -- Select all sites ST_Dwithin 1km of destination coordinates
    -- No need to worry about entrances
    TRUNCATE TABLE MENE.NearSites;
    numsites := 0;

    -- Parks:
    INSERT INTO MENE.NearSites (spid, type, supertype, prox, disthome, urbanpct, coastprox,
areagrid,
                lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal,
lc_wood_pasture, lc_sports_pitches, lc_golf,
                lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
                dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr,
dg_ramsar, poi_playground)
    SELECT spid, type, supertype,
         (1-(ST_Distance(geom, destination.geom_dest)/proximity)) AS prox,
         ST_Distance(geom, destination.geom_home) AS disthome,
         urbanpct, coastprox, areagrid,
      lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal, lc_wood_pasture,
lc_sports_pitches, lc_golf,
      lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
```

```
     dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr, dg_ramsar,
poi_playground
    FROM parks.parks_england
    WHERE ST_Dwithin(geom, destination.geom_dest, proximity) AND supertype IS NOT NULL;
    GET DIAGNOSTICS numrows = ROW_COUNT;
    numsites = numsites + COALESCE(numrows,0);

    -- Paths:
    INSERT INTO MENE.NearSites (spid, type, supertype, prox, disthome, urbanpct, coastprox,
areagrid,
                lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal,
lc_wood_pasture, lc_sports_pitches, lc_golf,
                lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
                dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr,
dg_ramsar, poi_playground)
    SELECT spid, type, supertype, (prox*0.95^(rank-1)) AS prox, disthome, urbanpct, coastprox,
areagrid,
      lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal, lc_wood_pasture,
lc_sports_pitches, lc_golf,
      lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
      dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr, dg_ramsar,
poi_playground
    -- ranks access points
    FROM (SELECT spid, type, supertype, prox, disthome, urbanpct, coastprox, rank() OVER
(PARTITION BY pid ORDER BY dist_acc ASC) AS rank, areagrid,
       lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal, lc_wood_pasture,
lc_sports_pitches, lc_golf,
       lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
       dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr, dg_ramsar,
poi_playground
    -- Find access points to close paths
      FROM (SELECT tbl1.pid, tbl1.spid, tbl1.type, tbl1.supertype,
                       (1-ST_Distance(tbl2.geom, destination.geom_dest)/proximity) AS prox,
                                      ST_Distance(tbl2.geom, destination.geom_home) AS
disthome,
                       tbl1.urbanpct, tbl1.coastprox, tbl1.areagrid,
        ST_Distance(tbl1.geom, destination.geom_dest) AS dist_acc,
        lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal, lc_wood_pasture,
lc_sports_pitches, lc_golf,
        lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
        dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr, dg_ramsar,
poi_playground
      FROM paths.paths_england AS tbl1 INNER JOIN
        -- Select paths that pass close to destination
        (SELECT DISTINCT pid, ST_ClosestPoint(geom_line, destination.geom_dest) AS geom
       FROM paths.paths
       WHERE ST_Dwithin(geom_line, destination.geom_dest, proximity)) AS tbl2
      ON tbl1.pid = tbl2.pid AND ST_DWithin(tbl1.geom, tbl2.geom, 10000)) AS tbl3) AS tbl4;
   -- Find nearest point on paths, then select all access points paths that are within 10km
   -- Rank access points on same pid and reweight prox score according to how close access
point is to destination
    GET DIAGNOSTICS numrows = ROW_COUNT;
    numsites = numsites + COALESCE(numrows,0);

    -- Beaches:
    INSERT INTO MENE.NearSites (spid, type, supertype, prox, disthome, urbanpct, coastprox,
areagrid,
                lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal,
lc_wood_pasture, lc_sports_pitches, lc_golf,
                lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
                dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr,
dg_ramsar, poi_playground)
    SELECT spid, type, supertype,
        (1-(ST_Distance(geom, destination.geom_dest)/proximity)) AS prox,
                ST_Distance(geom, destination.geom_home) AS disthome,
        urbanpct, coastprox, areagrid,
      lc_woods, lc_agrculture, lc_moors_heath, lc_mountain, lc_coastal, lc_wood_pasture,
lc_sports_pitches, lc_golf,
      lc_allotments, lc_seaside, lc_estuary, lc_rivers_canals, lc_lakes_reservoirs,
      dg_ancient_woodland, dg_sssi, dg_CPark, dg_natura2000, dg_nnr, dg_lnr, dg_ramsar,
poi_playground
    FROM beaches.beaches_england
    WHERE ST_Dwithin(geom, destination.geom_dest, proximity);
    GET DIAGNOSTICS numrows = ROW_COUNT;
    numsites = numsites + COALESCE(numrows,0);
    -- RAISE NOTICE ' Number sites near %: %', destination.visitid, numsites;
```

```
    IF numsites = 0 THEN
       -- MATCHCODE -1
       IF NOT TEST THEN
                      EXECUTE 'UPDATE mene.visit_match SET matchcode = -1 WHERE visitid =
'||destination.visitid;
              END IF;
    ELSE
        IF NOT TEST THEN
               EXECUTE 'UPDATE mene.visit_match SET matchcode = 0 WHERE visitid =
'||destination.visitid;
        END IF;

    -- (2) LOCATION SCORE
    -- -------------------
    -- 50 pts for proximity
    -- Linear Distance Decay: 50 * (1 - ST_Distance/proximity)
    UPDATE MENE.NearSites SET near_score = (50 * prox);

        -- Check for compatibility with stated travel distance
    UPDATE MENE.NearSites
    SET near_score = near_score * disthome/(destination.travdistlo*.5)
    WHERE disthome < destination.travdistlo*.5;

        UPDATE MENE.NearSites
    SET near_score = near_score * (destination.travdisthi*1.5)/disthome
    WHERE disthome > destination.travdisthi*1.5;

    -- (3) ENVIRONS SCORE
    -- ------------------
    -- urbanpct, rural, coastal, inland
    CASE destination.q2
      WHEN 'In a town or city' THEN
      UPDATE MENE.NearSites SET loc_score = (10*urbanpct     + 10*greatest(0,(coastprox-
coastthrshld)/(coastdist-coastthrshld)));
      WHEN 'In the countryside (including areas around towns and cities)' THEN
      UPDATE MENE.NearSites SET loc_score = (10*(1-urbanpct) + 10*greatest(0,(coastprox-
coastthrshld)/(coastdist-coastthrshld)));
      WHEN 'In a seaside resort or town' THEN
      UPDATE MENE.NearSites SET loc_score = (10*urbanpct     + 10*(1-greatest(0,(coastprox-
coastthrshld)/(coastdist-coastthrshld))));
      WHEN 'Other seaside coastline (including beaches and cliffs)' THEN
      UPDATE MENE.NearSites SET loc_score = (10*(1-urbanpct) + 10*(1-greatest(0,(coastprox-
coastthrshld)/(coastdist-coastthrshld))));
      ELSE
    END CASE;

    CASE destination.q5_02 -- Farmland
      WHEN 'Yes' THEN -- rural
      UPDATE MENE.NearSites SET loc_score = loc_score + (5*(1-urbanpct));
      ELSE
    END CASE;

    CASE destination.q5_03 -- Mountain, Wood or Moorland
      WHEN 'Yes' THEN -- rural
      UPDATE MENE.NearSites SET loc_score = loc_score + (5*(1-urbanpct));
      ELSE
    END CASE;

    CASE destination.q5_08 -- Another open space in the countryside
      WHEN 'Yes' THEN -- rural
      UPDATE MENE.NearSites SET loc_score = loc_score + (5*(1-urbanpct));
      ELSE
    END CASE;

    CASE destination.q5_09 -- A park in a town or city
      WHEN 'Yes' THEN -- urban
      UPDATE MENE.NearSites SET loc_score = loc_score + (5*urbanpct);
      ELSE
    END CASE;

    CASE destination.q5_13 -- Another open space in a town or city
      WHEN 'Yes' THEN -- urban
      UPDATE MENE.NearSites SET loc_score = loc_score + (5*urbanpct);
      ELSE
    END CASE;

    CASE destination.q5_14 -- A beach
```

```
     WHEN 'Yes' THEN -- coastal
     UPDATE MENE.NearSites SET loc_score = loc_score + (5*coastprox/coastdist);
     ELSE
    END CASE;

    CASE destination.q5_15 -- Other coastline
     WHEN 'Yes' THEN -- coastal
     UPDATE MENE.NearSites SET loc_score = loc_score + (5*coastprox/coastdist);
     ELSE
    END CASE;

    CASE destination.q4_13 -- Visits to a beach (sunbathing or paddling in the sea)
     WHEN 'Yes' THEN -- coastal
     UPDATE MENE.NearSites SET loc_score = loc_score + (5*coastprox/coastdist);
     ELSE
    END CASE;


    -- (4) TYPE SCORE
    -- --------------
    -- path    park   beach   country park   allotment golf
    UPDATE MENE.NearSites SET type_score = 0;

    -- Allotment
    -- ---------
    CASE destination.q5_10 -- An allotment or community garden
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 50 WHERE supertype = 'allotment';
     WHEN 'No' THEN
     UPDATE MENE.NearSites SET type_score = type_score - 10 WHERE supertype = 'allotment';
     ELSE
    END CASE;

    -- Country Park
    -- ------------
    CASE destination.q5_07 -- Country parks
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 25 WHERE supertype = 'country_park';
     WHEN 'No' THEN
     UPDATE MENE.NearSites SET type_score = type_score - 5 WHERE supertype = 'country_park';
     ELSE
    END CASE;

    -- Cemetery
    -- --------
    CASE destination.q5_08 -- Another open place in the countryside
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 3 WHERE supertype = 'cemetery';
     ELSE
    END CASE;
    CASE destination.q5_13 -- Another open place in a town or city
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 3 WHERE supertype = 'cemetery';
     ELSE
    END CASE;

    -- Park
    -- ----
    CASE destination.q5_09 -- A park in a town or city
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 10 WHERE supertype = 'park';
--              WHEN 'No' THEN
--              UPDATE MENE.NearSites SET type_score = type_score - 2 WHERE supertype =
'park';
     ELSE
    END CASE;
    CASE destination.q5_12 -- A playing field or other recreation area
     WHEN 'Yes' THEN
     UPDATE MENE.NearSites SET type_score = type_score + 8 WHERE supertype = 'park';
--              WHEN 'No' THEN
--              UPDATE MENE.NearSites SET type_score = type_score - 2 WHERE supertype =
'park';
     ELSE
    END CASE;

    -- Beach
    -- -----
```

```
    CASE destination.q5_14 -- A beach
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET type_score = type_score + 25 WHERE type = 'beach';
      WHEN 'No' THEN
      UPDATE MENE.NearSites SET type_score = type_score - 5 WHERE type = 'beach';
      ELSE
    END CASE;
    CASE destination.q4_13 -- Visits to a beach (sunbathing or paddling in the sea)
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET type_score = type_score + 25 WHERE type = 'beach';
      WHEN 'No' THEN
      UPDATE MENE.NearSites SET type_score = type_score - 5 WHERE type = 'beach';
      ELSE
    END CASE;

    -- Golf
    -- -----
    CASE destination.q4_19 -- Informal games and sport
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET type_score = type_score + 10 WHERE supertype = 'golf';
      WHEN 'No' THEN
      UPDATE MENE.NearSites SET type_score = type_score - 10 WHERE supertype = 'golf';
      ELSE
    END CASE;

    -- Path
    -- ----
    CASE destination.q5_06 -- A path
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET type_score = type_score + 10 WHERE type = 'path';
      ELSE
    END CASE;

    -- Nature
    -- ------
    CASE destination.q4_18 -- Wildlife Watching
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET type_score = type_score + 10 WHERE type = 'nature';
      ELSE
    END CASE;


    -- (5) LANDCOVER SCORE
    -- -------------------
    -- woods, water, farmland, playground, sports, nature, leisure, country park
    UPDATE MENE.NearSites SET lc_score = 0;

    -- Woods
    -- -----
    CASE destination.q5_01 -- A woodland or forest
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + 5 WHERE type = 'woods';
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_woods/areagrid) +
(5*lc_wood_pasture/areagrid) + (5*dg_ancient_woodland/areagrid);
      ELSE
    END CASE;

    -- Farmland
    -- --------
    CASE destination.q5_02 -- Farmland
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (15*lc_agrculture/areagrid);
      ELSE
    END CASE;

    -- Mountain, Hill, Moorland
    -- ------------------------
    CASE destination.q5_03 -- Mountain, Hill, Moorland
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_moors_heath/areagrid) +
(10*lc_mountain/areagrid);
      ELSE
    END CASE;

    -- River, Canal, Lake or Reservoir
    -- ------------------------------
    CASE destination.q5_04 -- River, Lake or Canal
```

```
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_rivers_canals/areagrid) +
(10*lc_lakes_reservoirs/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_03 -- Fishing
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_rivers_canals/areagrid) +
(10*lc_lakes_reservoirs/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_12 -- Swimming Outdoors
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (5*lc_rivers_canals/areagrid) +
(5*lc_lakes_reservoirs/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_17 -- Watersports
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (5*lc_rivers_canals/areagrid) +
(10*lc_lakes_reservoirs/areagrid);
      ELSE
    END CASE;


    -- Seaside or Esturary
    -- ------------------
    CASE destination.q4_03 -- Fishing
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_seaside/areagrid) +
(10*lc_estuary/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_12 -- Swimming Outdoors
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_seaside/areagrid) +
(10*lc_estuary/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_17 -- Watersports
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_seaside/areagrid) +
(10*lc_estuary/areagrid);
      ELSE
    END CASE;

    -- Sports Pitches
    -- --------------
    CASE destination.q5_12 -- Playing Fields or Other Recreation Area
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_sports_pitches/areagrid);
      ELSE
    END CASE;
    CASE destination.q4_19 -- Informal games and sport
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*lc_sports_pitches/areagrid)+
(10*lc_golf/areagrid);
      ELSE
    END CASE;

    -- Nature
    -- ------
    CASE destination.q4_18 -- Wildlife watching
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (5*greatest(dg_nnr, dg_lnr,
dg_natura2000, dg_sssi, dg_ramsar)/areagrid);
      ELSE
    END CASE;

    -- Playgrounds
    -- -----------
    CASE destination.q5_11 -- A children's playground
      WHEN 'Yes' THEN
      UPDATE MENE.NearSites SET lc_score = lc_score + (10*poi_playground);
      ELSE
    END CASE;
```

```
    -- (6) RECORD BEST 3 MATCHES
    -- -------------------------
    UPDATE MENE.NearSites SET score = near_score + COALESCE(loc_score,0) +
COALESCE(type_score,0) + COALESCE(lc_score,0);
    IF NOT TEST THEN
        matchcnt := 1;
      FOR matches IN SELECT spid, near_score, loc_score, score FROM MENE.NearSites ORDER BY
score DESC LIMIT 3 LOOP
        EXIT WHEN NOT FOUND;
        EXECUTE 'UPDATE mene.visit_match SET match'||matchcnt||'id = '||matches.spid||',
match'||matchcnt||'score = '||matches.score||' WHERE visitid = '||destination.visitid;
        matchcnt := matchcnt + 1;
      END LOOP;
    END IF;
   END IF;
  END IF;

   IF pcounter = nprint THEN
    nprocessed := (SELECT count(*) FROM mene.visit_match WHERE matchcode IS NOT NULL);
    RAISE NOTICE '     Visits processed: % (% of % = % pct done)', counter, nprocessed,
ntoprocess, (nprocessed/ntoprocess);
    pcounter := 0;
   END IF;

   pcounter := pcounter + 1;
   counter := counter + 1;

 END LOOP;

 IF NOT TEST THEN
   DROP TABLE IF EXISTS MENE.NearSites;
 END IF;

END;
$$ LANGUAGE plpgsql;
```