Risk of sustained ignition mapping for the Peak District National Park

Joint funded by:
Peak District Fire Operations Group and MoorLIFE 2020

Peak District F.O.G Fire Operations Group

Life

NATURA 2000

Prepared by:

MOORS FOR THE FUTURE PARTNERSHIP

Moors for the Future Partnership

2019

Members of the strategic management group of the Moors For the Future Partnership

Environment Agency
Peak District National Park Authority
Natural England
Yorkshire Water Services
Severn Trent Water
Unites Utilities
Royal Society for the Protection of Birds
Moorland Association
National Trust
National Farmers Union
Pennine Prospects

Members of the Peak District Fire Operations Group

Peak District National Park Authority
National Trust
Cheshire Fire & Rescue Service
Derbyshire Fire & Rescue Service
Greater Manchester Fire & Rescue Service
South Yorkshire Fire & Rescue Service
Staffordshire Fire & Rescue Service
West Yorkshire Fire & Rescue Service
Severn Trent Water
United Utilities
Yorkshire Water
Royal Society for the Protection of Birds
Chatsworth
Natural England
Staffordshire Wildlife Trust
Eastern Moors Partnership
Sheffield Wildlife Trust

Prepared by

Moors for the Future Partnership
The Moorland Centre, Edale, Hope Valley, Derbyshire, S33 7ZA, UK

T: 01629 816291
M: 07980 755778
E: sam.dixon@peakdistrict.gov.uk
W: www.moorsforthefuture.org.uk

Suggested citation: Dixon, S.G. and Chandler, D. (2019). Producing a risk of sustained ignition map for the Peak District National Park Moors for the Future Report, Edale.

Contents

List of Abbreviations

AUC	Area Under Curve
BIC	Bayesian Information Criterion
CEH	Centre for Ecology and Hydrology
ESA	Environmentally Sensitive Areas
FRS	Fire and Rescue Services
IMD	Index of Multiple Deprivation
IRS	Incident Recording System
LCM	Land Cover Map
MCE	Multi Criteria Evaluation
MFFP	Moors For the Future Partnership
PDNPA	Peak District National Park Authority
ROC	Receiver Operating Curve
SAC	Special Area of Conservation

List of Figures

List of Tables

1. Introduction

The Peak District FOG group currently use the ignition risk map produced by McMorrow and Lindley (2006) to guide their work. Since the publication of that report, Moors for the Future Partnership (MFFP) have collated a database of wildfire locations in the Peak District and South Pennine moors (Titterton and Crouch, 2018). The database is both more complete, with data procured from both National Park rangers, Fire Rescue Services and local land owners; and up to date, extending from 1976 (with limited data in the early years) until the present day. Furthermore, significant changes in land cover have occurred following MFFP's restoration work since 2003 possibly influencing risk of ignition. Therefore, it is timely to update the ignition risk map.

It should be noted the maps produced both here and in McMorrow and Lindley (2006) only show 'risk of reported ignition', i.e. that a wildfire has started, been reported and either the Fire Rescue Services or National Park rangers have attended. In other words, the ignition risk maps show only the likelihood of ignition and give no indication of severity (the amount of damage or harm an ignition could create). The ignition risk maps could be used alongside the Fire Severity Index (FSI) produced by the Met Office, which is an assessment of how severe a fire could become if one were to start but do not represent the risk of wildfires occurring. The report treats all wildfire ignitions equally with no stratification by size or time of year, for example. As the wildfire database increases in size and detail, such stratified approaches may become possible.

This report builds upon McMorrow and Lindley's (2006) work for the Peak district National Park only. A later analysis was undertaken for the whole of the South Pennines SAC but the lack of accurate fire occurrence and predictor data impeded the production of a spatial risk map. This lack of data still exists to a lesser extent today, hence the limiting of this analysis to the Peak District only. However, an extension of the approach to the South and West Pennine moors should be a priority. Currently, this is limited by a lack of wayline data and the spatial inaccuracy of wildfire occurrence data.

Aim

- Update the ignition risk map produced by McMorrow and Lindley (2006) using up to date information

Objectives

- Test the performance of McMorrow and Lindley's (2006) risk map for recent wildfires
- Update the ignition risk map of McMorrow and Lindley (2006) using up to date wildfire information and predictor datasets
- Pilot a logistic regression approach for producing the ignition risk map

2. Data and study area

2.1. Wildfire Database

A limiting factor in the previous report was the locational accuracy and completeness of the wildfire record. MFFP has since created a wildfire database to help address these problems. The wildfire database is a record of wildfires from 1976 to the present. The first four decades of the database include fires from the Peak District National Park Authority (PDNPA) rangers only, but from 2009 onwards fires recorded by Fire Rescue Services through their Incident Recording System (IRS) and large landowners (e.g. National Trust) are included. Hence, the database can be considered a comprehensive record of wildfire occurrence since 2009. However, the locational accuracy of wildfires is still an issue. It is not known whether the location recorded is the point of ignition, the centre of the burnt area or the location of the fire engine. Problematically, it is likely that locations in the database are a mix of all three. The recommendation of McMorrow and Lindley (2006) is reiterated here that location collection should be standardised, for example fire perimeter or the estimated point of ignition.

2.2. Predictor datasets

Where possible, recent data were procured to replace the versions used in the previous report. In some cases like-for-like replacements were possible, in others pragmatic replacements were chosen and in few cases no recent data were available. All predictor datasets have a spatial resolution of 50m. An overview of each dataset is provided in Table 1 and detailed information on each dataset is available in Annex 1: Data .

Table 1: Comparison of datasets used by McMorrow and Lindley (2006) and those used here

Feature	Dataset used by McMorrow and Lindley (2006)	Dataset used here
Habitat	PDNP moorland habitat map based on Environmentally Sensitive Areas (ESA) habitat map	Centre for Ecology and Hydrology's (CEH) Land cover Map (LCM) 2015
Major and Minor Roads	Unknown	OS Open Roads
Public Rights of Way (PROW)	Unknown	Local Highway Authority derived data held by the PDNPA
Waylines[1]	MFFP dataset comprising digitised waylines from aerial imagery	As in McMorrow and Lindley (2006)
Pennine Way (high popularity sections)	Identified from August 2004 and January 2005 surveys undertaken by PDNPA	As in McMorrow and Lindley (2006)

[1] Waylines were mapped by MFFP during a previous project in 2005 and are defined as tracks visible on aerial imagery that are not PROW.

Settlements	Urban centres generated from Enumeration Districts and based on 1991 Census	Index of Multiple Deprivation (IMD) deciles 1-3 (i.e. lowest 30%)
Lay-bys	N/A	OS MasterMap cartographic text
Car parks	N/A	OS MasterMap cartographic text

2.3. Study Area

The study area covers Section 3 moorland within the Peak District National Park (Figure 1).

Figure 1: Ignition risk map boundary based on Section 3 Moorland within the National Park

3. Approach 1: Multi Criteria Evaluation (MCE)

3.1. Testing the previous model using updated fire database

McMorrow and Lindley (2006) used two measures to test model performance, the mean risk score of test fires and a Mann-Whitney test between training and test fire risk scores. The Mann-Whitney could not be used as there was no record of the specific fires used for training and testing in the 2006 report. Therefore, we relied upon using the mean risk score of fires from 2009-2018 compared with those in Table 6.8 of McMorrow and Lindley (2006).

3.2. Inspecting the wildfire database

First, a visual inspection of fire distribution was undertaken. A clear shift was identified in the spatial distribution of pre-2009 fires compared with post 2009 fires (Figure 2). Changes in the data collection methods through time could be causing this shift. Pre-2008, only PDNP ranger data is included in the database whereas both PDNP and FRS data is included from 2009 onwards. Therefore, it is possible the relatively large number of fires recorded on Kinder and Bleaklow pre-2009 is in part driven by over representation, as they are focus areas of the rangers. However, the lack of recorded fires from 2009 onwards suggests wild fire risk in these areas has reduced. Due to the change in wildfire distribution, new models were built using data from 2009 onwards only. Whilst this does reduce the sample size, it is preferable to using the non-stationary full dataset that would violate statistical assumptions.

3.3. Updating the Multi Criteria Evaluation (MCE) models

The MCE analysis was updated for the two highest performing models in McMorrow and Lindley (2006), models 1c and 2g. It was decided that due to time and budget constraints, weightings for the MCE analysis would be kept the same as those in the previous analysis. It is important to note that these weights are likely not wholly appropriate and model results would be more robust if these weightings were revised. For example, high popularity Pennine Way sections are given an 18% weighting in Model 1c, yet show no distance decay relationship to wildfires between 2009-2018 (Annex 2: Detailed Methodologies). Updating the weightings would involve stakeholder engagement, likely in the form of a workshop and/or questionnaire as per McMorrow and Lindley (2006).

The scored distance decay rasters for each predictor were updated by producing histograms as in McMorrow and Lindley (2006). These give the frequency of wildfires within 'x' metre distance classes from each of the human factors included in the models. An example map is provided for minor roads in Figure 3 and its associated histogram is shown in Figure 9. The histograms produced and the resulting scored distance classes are provided for each dataset in Annex 2: Detailed Methodologies.

The Centre for Ecology and Hydrology (CEH) Land Cover Map (LCM) habitat dataset was scored using the same procedure McMorrow and Lindley (2006) used to score the

Environmentally Sensitive Areas (ESA) habitat classes. An area weighted risk score was calculated for each of the 21 LCM classes. This gives each class a score representing the number of fires per unit area in the study area. Scores attributed to each LCM 2015 class are provided in Annex 1: Data. The two models were run using the raster calculator in ArcGIS. The equations used to build the models were identical to those used in McMorrow and Lindley (2006), as the weightings remain the same. The full equations used are provided in Annex 2: Detailed Methodologies.

3.4. MCE model results

Mean risk score according to models produced by McMorrow and Lindley (2006) at known fire locations have reduced since 2009 (Table 2). Model 1c saw a greater reduction of -0.77 compared with the -0.69 reduction seen when using model 2g. McMorrow and Lindley's (2006) two best performing models were updated using wildfires since 2009 according to MFFP's wildfire database. Scores for each distance decay layer were updated, but the weightings for each layer within the model were maintained from McMorrow and Lindley (2006). Previous and updated risk maps for model 1c and model 2g are provided in Figure 4 and Figure 5 respectively. Both models show a clear change in the distribution of high risk from high elevation regions (principally Kinder and Bleaklow surrounding the Pennine Way) to the periphery, particularly in areas of rural-urban interface. This is driven by the change in wildfire distribution according to the wildfire database (Figure 2) and the associated change in scores attributed to the distance decay layers (Annex 2: Detailed Methodologies). Risk areas are also less fragmented, with greater areas of continuous low/medium risk.

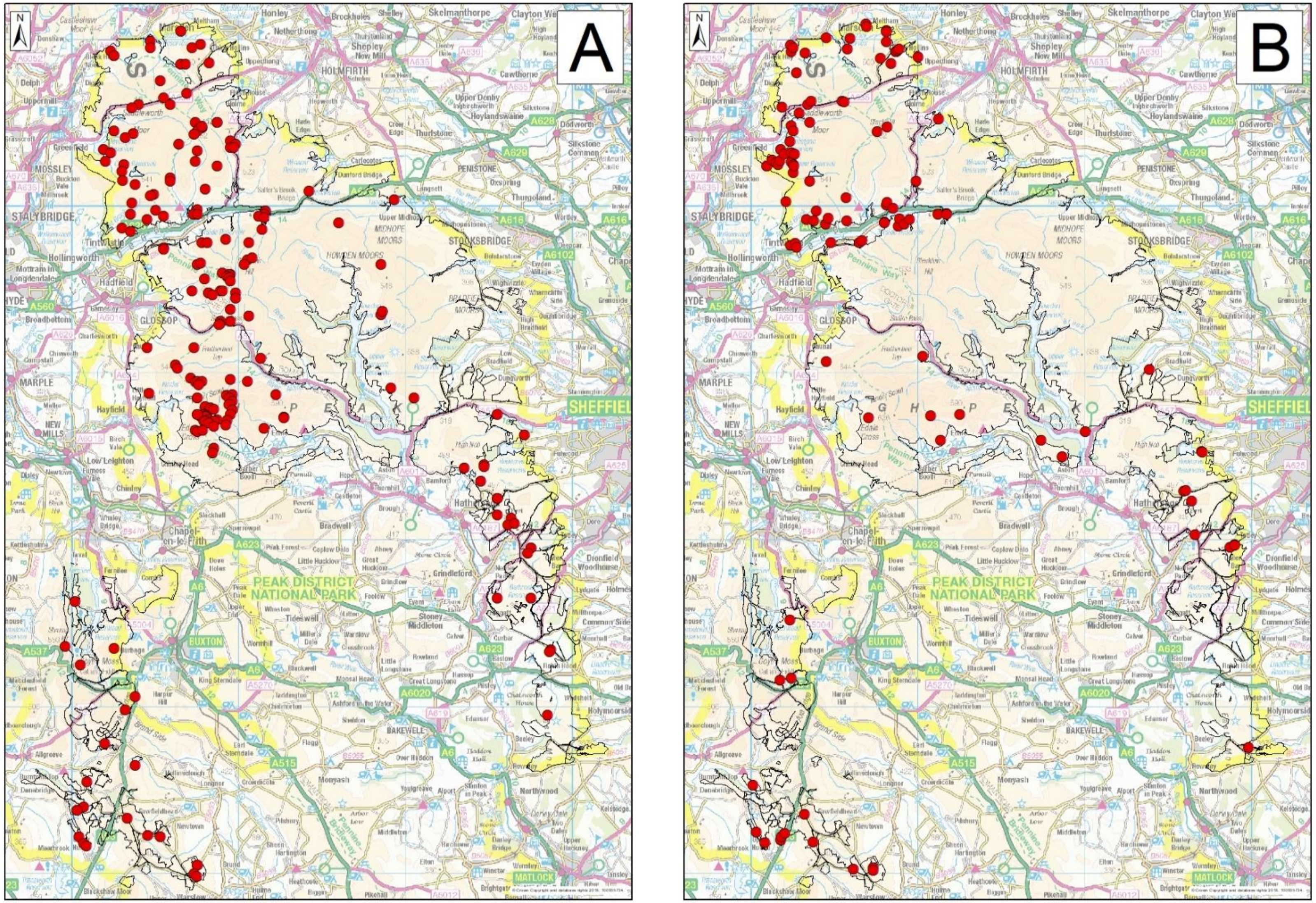

Figure 2: Wildfire locations during the years 1976-2003 (A) and 2009-2018 (B). Data from MFFP's wildfire database (Titterton and Crouch, 2018).

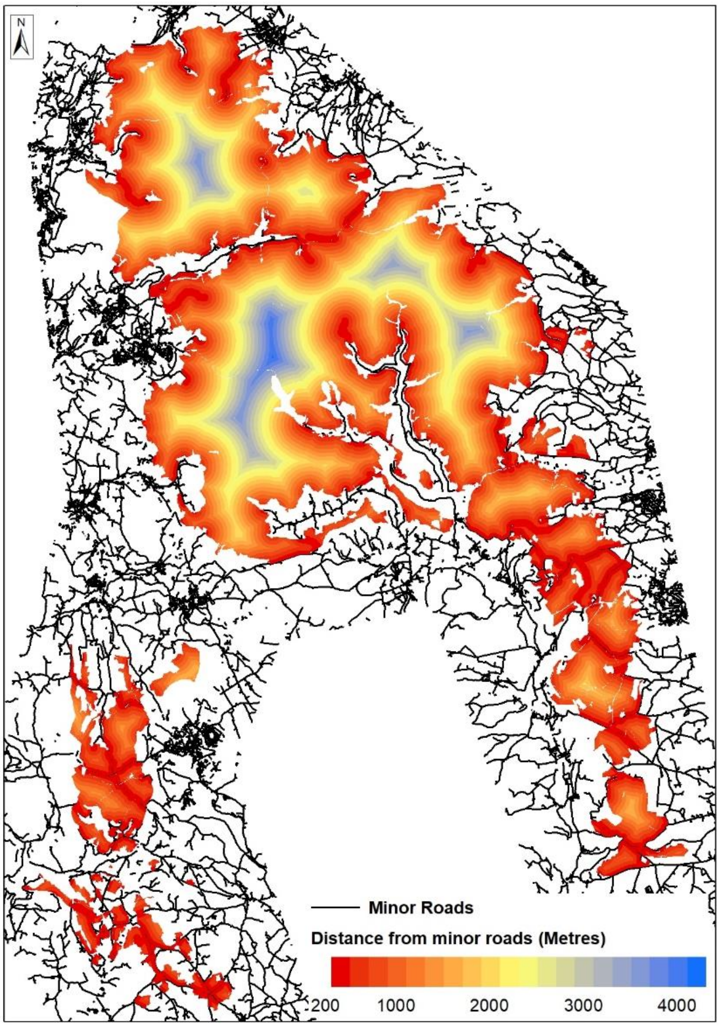

Figure 3: Example distance decay map from which the MCE histograms were produced and scores assigned.

Table 2: Mean ignition risk score from models 1c and 2g at wildfire locations

	Wildfires occurring between 1976-2003 (from McMorrow and Lindley, 2006)	Testing on wildfires occurring between 2009-2018	Change
Model 1c	5.39	4.62	- 0.77
Model 2g	6.64	5.95	- 0.69

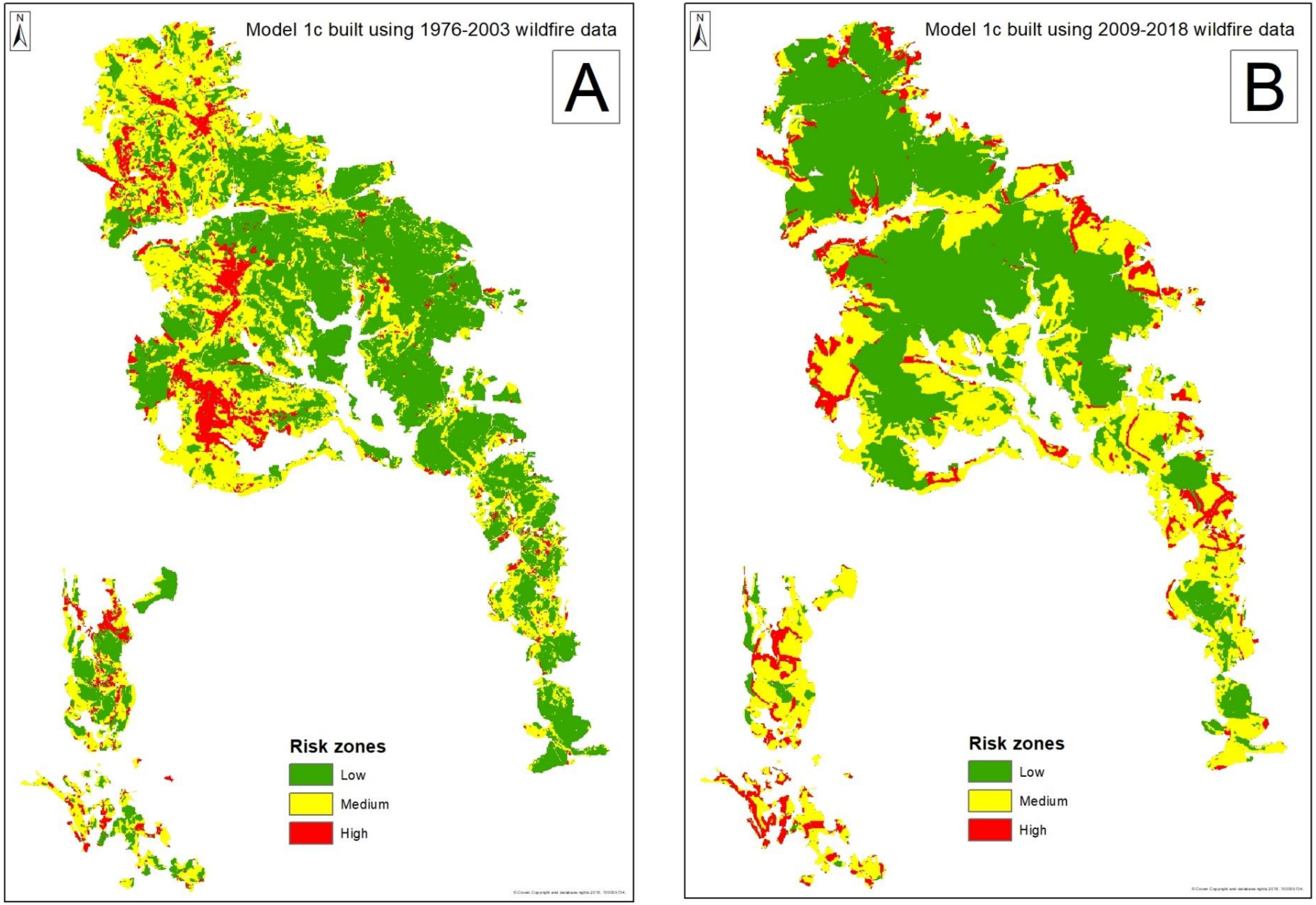

Figure 4: Previous MCE derived model 1c risk of ignition map (A) compared with updated risk map (B)

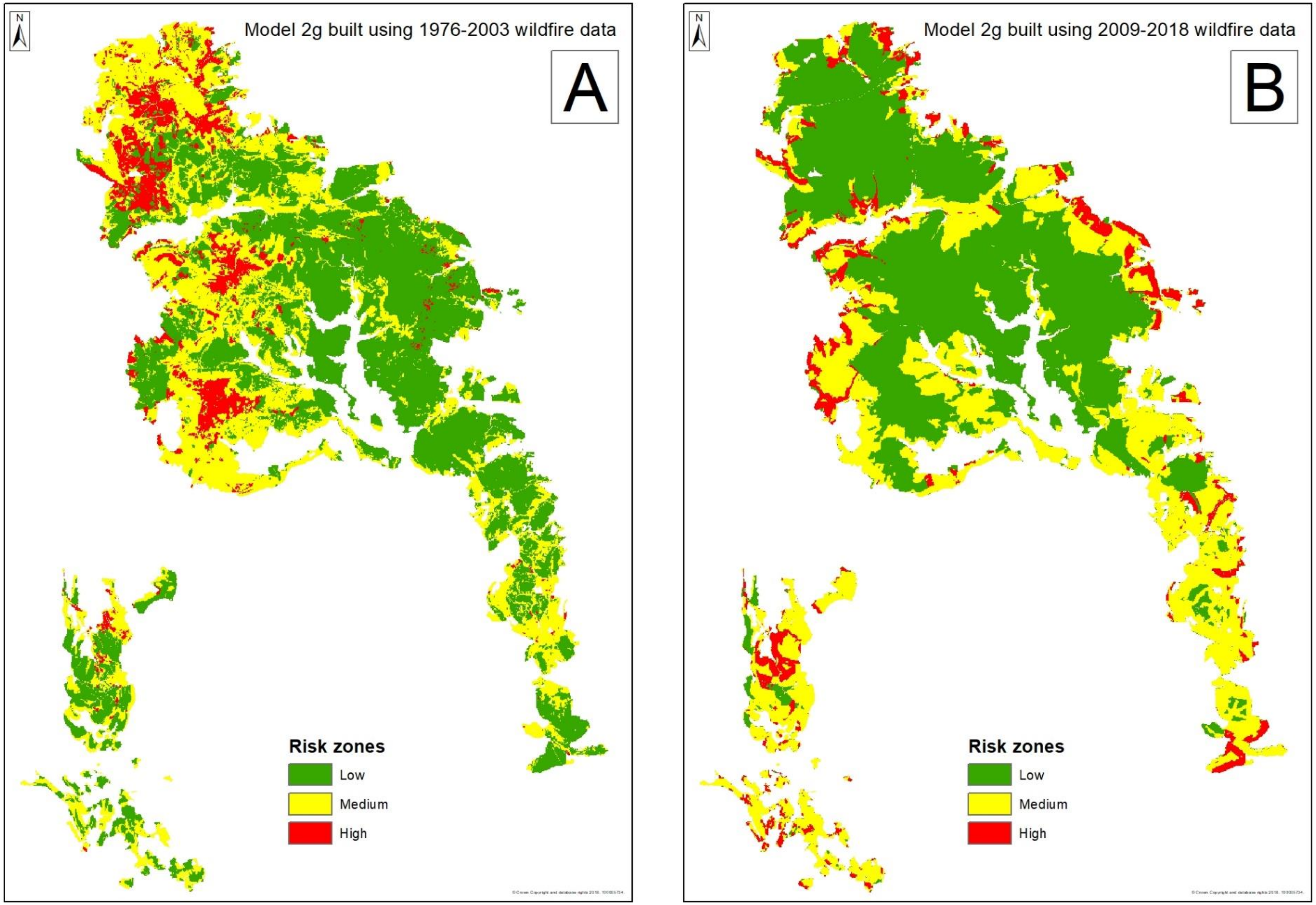

Figure 5: Previous MCE derived model 2g risk of ignition map (A) compared with updated risk map (B)

4. Approach 2: Logistic Regression

Several recent studies have utilised a logistic regression approach to mapping wildfire risk (e.g. Martinez et al., 2009 and Catry et al., 2009). Logistic regression has the benefit of being objective compared with the subjective weighting of the MCE approach. Compared with MCE, logistic regression models can be more rigorously tested (for example using a ROC curve) providing greater insight into performance. However, logistic regression is only as good as the data used. Therefore, the relatively small sample size available post 2009 (111 wild fires recorded) and the uncertainty regarding wildfire location accuracy are potential issues with this approach. Models built using a statistical approach require 'sense checking', ensuring that the variables selected make physical sense and are not simply statistical artefacts.

4.1. Data preparation

Data preparation was undertaken in ArcMap. Distance decay rasters were not scored, as in the MCE, but left in their raw form as distance (in metres) from features. The habitat layer was identical to that used in the MCE approach, with an area-weighted score between 0-10 given to each habitat (see section 3.3). The model requires non-ignition sample points as well as the ignition sample points (wildfires recorded 2009-2018) for training. These non-ignition points were selected randomly from the study area, excluding areas within 200m of a known ignition point. The 200m buffer aims to account for locational uncertainty of recorded wildfires, to ensure no ignition point was accidently included as a non-ignition point. Twice as many non-ignition points (222) than ignition points (111) were created, to better represent the greater variability than found in the ignition sample (Catry et al.,2009). We then extracted values from each of the predictor datasets at each ignition and non-ignition sample point. These extracted values became the dataset for building the model. We used a 60% random sample of this dataset for training the model, leaving 40% for independent testing.

4.2. Building the logistic regression model

The logistic regression model was built using the programming language R. The full code used is provided in Annex 4: R code used to build and assess logistic regression models. The model was built using a backwards stepwise procedure, as to produce a simpler model therefore reducing the likelihood of overfitting (See Annex 2 (9.2.7) for more information). As the training data sample size is relatively small (222 points), selecting model variables using just one sample may produce an anomalous result. Therefore, 10 models were run using different samples and the final model variables were selected using the most commonly included variables. The results of this procedure are in Table 3. The final model variables selected were distance from IMD deciles 1-3, minor roads and waylines. Whilst waylines were only selected in 7/10 models, public rights of way were selected in the remaining three. Therefore, it seems the model requires information on foot access routes and so waylines were included. Lay-bys were not included, even though 7/10 models selected them, to avoid overfitting. Interestingly, the logistical regression procedure did not include car parks in any model. It is possible that the car park dataset used (identified using

OS MasterMap labels) does not capture all car parks in the study area, limiting its predictive capacity.

Table 3: Variables included in each of the 10 model runs

	Model run										
Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Sum**
IMD	X	X	X	X	X	X	X	X	X	X	10
Lay-by	X	X				X	X	X	X	X	7
Major				X	X						2
Minor	X		X	X	X	X	X	X	X	X	9
Wayline	X	X		X	X		X	X	X		7
PWHiPop	X					X		X	X	X	5
PROW			X			X	X			X	4
Car Park											0
Habitat											0

The variable coefficients were estimated by running the final model selection on 10 random samples, then averaging the coefficients to provide the final model. Table 4 shows the final variables chosen and their coefficients. The Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) was used to assess model performance.

Table 4: Selected logistic regression model variables and their relative coefficients

	Coefficient	**Standard deviation**	**Standard deviation as % of mean**
Intercept	1.906	0.316	17
IMD deciles 1-3	-0.0002889	0.000041	14
Distance to minor roads	-0.0008911	0.000133	15
Distance to waylines	-0.0049403	0.000840	17

4.3. Logistic regression model results

The final logistic regression model was produced by running the selected model (Table 4) on 10 different samples, then averaging the coefficients to provide the final model. Figure 6 shows the ignition risk map produced using this final model. The final model had an average AUC score of 0.83. More detailed performance analysis of logistic regression model is provided in Annex 3: Logistic regression performance metrics.

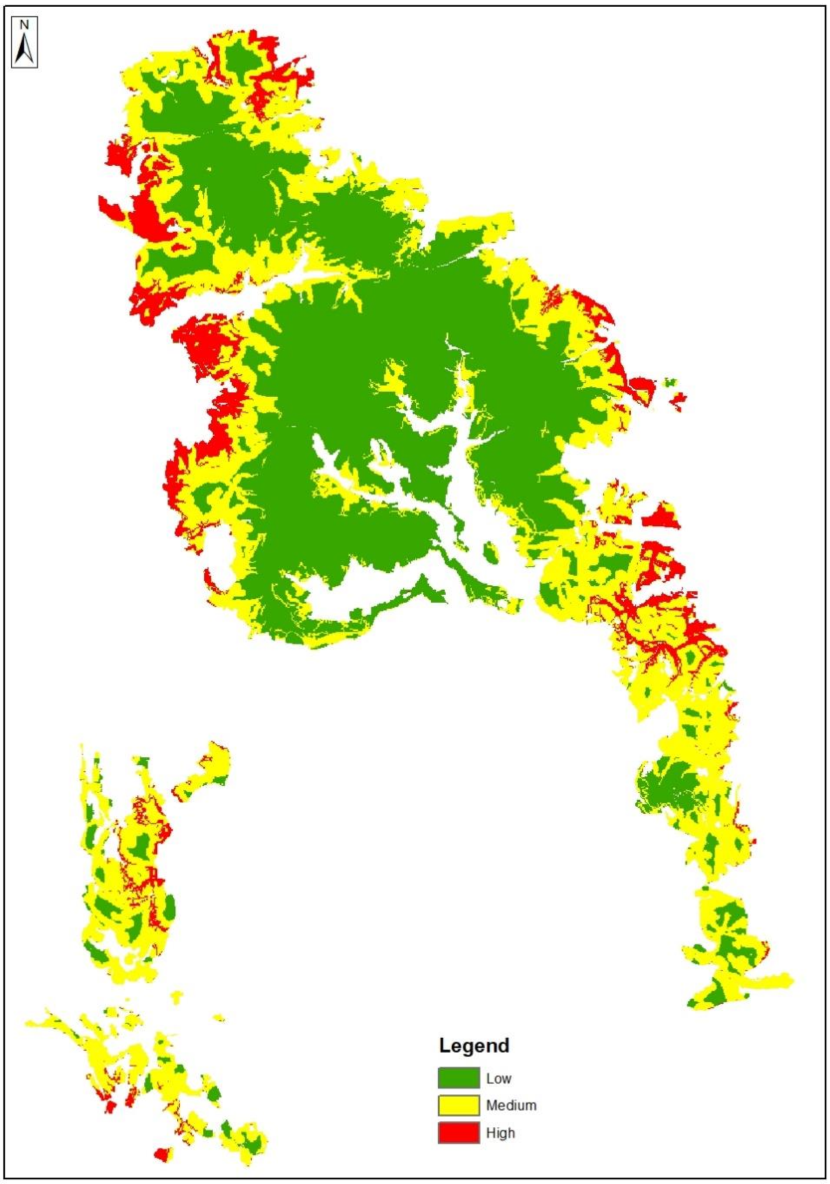

Figure 6: Ignition risk map produced using logistic regression model using fires during the period 2009-2018

5. Discussion

5.1. MCE approach

The reduction in mean ignition risk score for 2009-2018 wildfire locations compared to those between 1976-2008 suggests there is a requirement to update the mapping (Table 2). That the mean risk score has reduced is not surprising when considering the difference in fire distribution between the two periods as shown in Figure 2. The change of wildfire distribution shown in Figure 2 indicates that high risk areas have shifted, with few fires on Kinder and Bleaklow, previously some of highest risk zones.

All ignition risk maps shown in the report use the 50th and 90th percentiles to define low/medium/high risk. This means that high risk areas are defined as the highest 10% of the risk score (MCE) or probability score (logistic regression). This is a relatively arbitrary threshold, and we recommend the users of the ignition risk maps think carefully about what they would define as high risk and customise the maps accordingly. The raw raster data files are available from MFFP upon request if a user wishes to produce a customised version using locally-defined thresholds and colour scale.

The updated risk of sustained ignition maps (Figure 4 and Figure 5) reflect the change in fire distribution shown in Figure 2. This change in risk distribution is driven by the updated distance decay layer scores (Annex 1: Data), as weightings applied to each distance decay layer in the MCE model have been kept constant. The updated risk areas appear less fragmented, with greater areas of continuous low/medium/high risk. The elimination of the distance from Pennine Way high popularity layer due to its lack of relationship to wildfires may be one cause of this. The more likely cause, however, is the change in habitat data used. The stakeholder engagement undertaken by McMorrow and Lindley (2006) resulted in habitat being given a weighting of 50% in both models. This 50% weighting means that in the MCE model, habitat is driving the ignition risk. This is in stark contrast to the logistic regression approach which will be explored later. The LCM2015 class 'Bog' covers a large percentage of the study area, and whilst 29 fires occurred within 'Bog' it had the fewest number of fires by density. Using the methodology from McMorrow and Lindley (2006), this means 'Bog' was attributed a score of one. Furthermore, the LCM2015 dataset does not have all classes desired. For example, it does not include Molinia grassland. This habitat scoring has clearly driven the final MCE risk maps, with low risk zones aligning with 'Bog' habitat scored one in the habitat layer (Figure 7). This also reveals the problem with maintaining the weightings as they were in McMorrow and Lindley (2006). Given the clear shift in wildfire distribution, it is likely that these weightings require updating. We would strongly recommend the user review the weightings prior to using the MCE maps for decision making.

5.2. Logistic regression approach

Analysing the variables chosen by the logistic regression model can help understand the drivers of risk of ignition in the study area. However, other variables not included in the model might be of greater importance. From the variables selected, access seems to be an important factor, with areas close to large population centres, minor roads and paths having greater risk of ignition (Table 3). Furthermore, whilst lay-bys were not included in the final selection they were selected from several samples, again reiterating the importance of access. Interestingly, habitat was not included in any model. This could be because the habitat data used is not accurate enough, was too coarse (50m), does not have the required number of classes or is simply not a significant factor in risk of ignition. This goes against the stakeholder weightings which gave habitat by far the greatest importance at 50%. The ignition risk maps only describe the risk of ignition, not wildfire severity/extent/danger regarding which habitat may be of greater importance. It can be difficult to disentangle the risk of ignition from resulting fire behaviour'. Therefore, it is possible that a misunderstanding during the stakeholder workshop causes the discrepancy in habitat importance between the MCE and logistic regression.

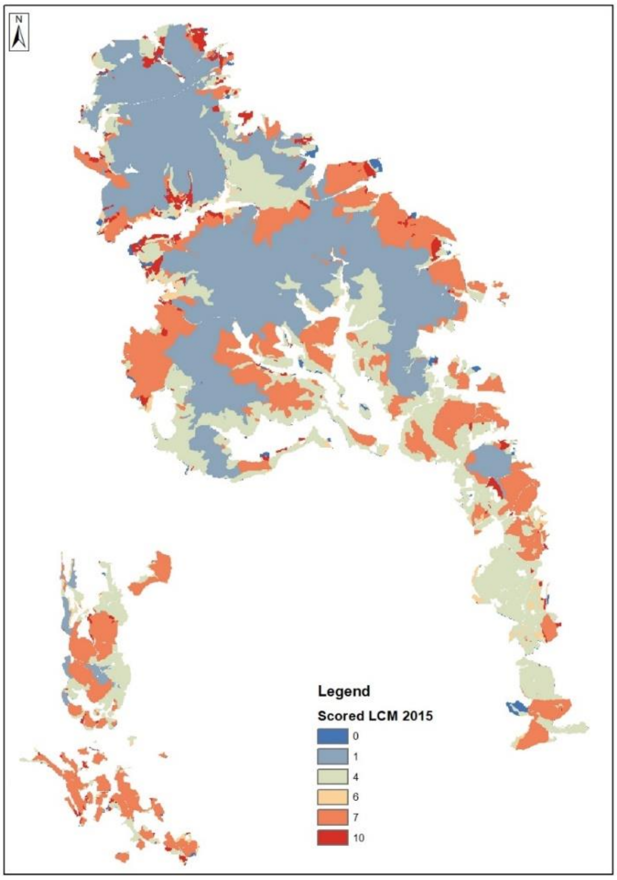

Figure 7: Scored LCM 2015 habitat map. Annex 1: Data includes a full list of LCM 2015 classes and their scores.

The performance metrics shown in Annex 3: Logistic regression performance metrics give an overview of the quality of the model. The average AUC of the 10 sample models of 0.83 is typically given a performance rating of 'good'. The model performs well regarding sensitivity (0.94), i.e. the avoidance of false negatives (predicting non ignition where actually there was an ignition). However, the models specificity is much lower (0.44), highlighting its tendency for predicting false positives (predicting ignition where actually the cell was non-ignition). This result is unsurprising when predicting wildfire ignitions, as the vast majority of the study area is covered by non-ignition with only a few cells showing ignitions. A 0.5 threshold (i.e. the model predicts 'fire' if there is a >50% chance of a fire occurring at this location) was used to calculate these statistics. The user should make this threshold decision, as the cost of a false positive is likely not equal to the cost of a false negative. For example, attending a call-out in which no ignition has occurred may be preferable to not attending a call-out in which there is an ignition. Therefore, it may be preferable to set a higher threshold (e.g. model predicts 'fire' if there is a >70% chance of a fire occurring at this location) to reduce the likelihood of a false negative (the model predicts no ignition when actually there is an ignition).

The resulting ignition risk map (Figure 6) shows high risk areas on the periphery of the study area, particularly surrounding cities/towns such as Sheffield in the east and Glossop and Mossley in the west. The IMD layer in the model likely drives this distribution. The areas surrounding the reservoirs in the Longdendale valley are only attributed medium risk, despite a number of wildfires occurring here since 2009. The high importance given to IMD in the model cannot capture the risk in this region, which might have been better captured by a greater weighting of the access layers. The logistic regression model is only as good as the data used within it and possibly an additional layer of visitor counts or known honeypot locations might help model this area.

5.3. Comparison between the MCE and logistic regression ignition risk maps

Both the MCE and Logistic regression maps show a similar pattern of risk of ignition. In general, the high elevation regions in the centre of the study region have low risk with high risk areas being confined to the periphery. However, some differences do exist. The MCE maps have greater medium risk areas around the Derwent reservoirs and the Edale valley compared to the logistic derived map which classes these areas as low risk. This difference is due to the MCE map being driven by habitat rather than distance from IMD deciles 1-3 as in the logistic regression. The logistic regression derived map seems to overstate the risk on the western edge of the study area around Glossop. This area has similar characteristics to Dovestones (to the north) but unlike Dovestones has seen few fires since 2009.

6. Summary

This report has updated the ignition risk map originally produced by McMorrow and Lindley (2006). We have utilised the wildfire database collated by Moors for the Future Partnership and up to date predictor datasets where available. Key findings from the report are listed below:

- Wildfire distribution has changed through time. Distribution of wildfires from 2009 onwards show clear differences to wildfires from 1976-2008.
- Both MCE and logistic regression derived maps show different ignition risk distribution across the study area compared with the best performing models from McMorrow and Lindley (2006).
- The scored distance decay rasters have been updated within the Multi Criteria Evaluation (MCE) model, however the weightings of layers remain the same as those within McMorrow and Lindley (2006). We recommend that stakeholders update these weightings if this model is used for decision-making.

The report has also highlighted several areas for further research. As the wildfire database continues to grow, fires could be stratified by size, time of year, antecedent conditions, cause of ignition, region etc., allowing more specific ignition risk maps to be produced. Since 2009, the wildfire database extends to the full South Pennines SAC. The spatial extent of the predictor datasets, in particular waylines, currently limits the risk map to the Peak district National Park. Extending this to cover the whole of the SAC would be a logical next step. Many of the recommendations made by McMorrow and Lindley (2006) are still valid today. In particular, standardising the collection of wildfire location data would facilitate improved modelling of ignition risk. Collecting the expected ignition point of the fire and the burn perimeter would be the ideal.

7. Acknowledgements

The authors would like to thank Julia McMorrow for her helpful advice on the methodologies and for her comments on the manuscript, which greatly improved the report.

8. References

Catry, F.X., Rego, F., Bação, F. and Moreira, F. (2009). Modeling and mapping wildfire ignition risk in Portugal. International Journal of Wildland Fire. 18(8), 921-931.

Martínez, J,. Vega-Garcia, C. and Chuvieco, E. (2009). Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management. 90(2), 1241-1252.

McMorrow, J. and Lindley, S. (2006). *Modelling the spatial risk of moorland wildfire.* Moors for the Future Partnership, Moorland Centre, Edale and GeoInformatics Research Group, Geography, School of Environment and Development, University of Manchester.

Titterton, P. and Crouch, T. (2018). *Updated Wildfire Database Report 2018: A guide to the method used in creation of the wildfire database and an analysis of trends associated with key variables.* Moors for the Future Partnership, Moorland Centre, Edale.

9. Annexes

9.1. Annex 1: Data

9.1.1. Habitat

The Centre for Ecology and Hydrology's (CEH) Land Cover Map (LCM) 2015 was used. This dataset covers the full study area at a practical resolution (25m) and is freely available to the national park under a DEFRA licence. However, it does not include a bare peat class, which was given high importance in the previous report. It would also be preferable to use pre-burn vegetation, such as the LCM2007 dataset, however this was not available under the DEFRA licence.

9.1.2. Major and Minor Roads

The OS Open Roads dataset was used to identify both major and minor roads. Major roads were classed as A roads (as in the previous report), with all other roads classed as minor.

9.1.3. Public Rights of Way (PRoW)

PRoW data were obtained from the PDNPA for the relevant highway authority that intersect the National Park.

9.1.4. Waylines

Waylines were mapped by MFFP during a previous project in 2005 and are defined as tracks visible on aerial imagery that are not PROW. No more recent data was available to update this dataset.

9.1.5. Pennine Way (High Popularity)

Pennine way popularity was mapped by McMorrow and Lindley (2006). No more recent data was available to update this dataset.

9.1.6. Settlements

It was deemed that no major new settlements have been built since the previous report and so the same dataset was used for the MCE. A new dataset was considered, however the simple relationship shown when using 5.5km wide distance bands was considered unlikely to change without drastic changes to the underlying settlement data.

Additional datasets were procured after an initial investigation of fire distribution. These were as follows.

9.1.7. Index of Multiple Deprivation (IMD)

After a visual investigation of fire distribution, it was clear that easily accessible locations close to large populations on the fringe of the park show high fire densities. Several methods could be used to describe the proximity to large urban areas, such as the method used to create settlements in the previous report (based on population). IMD was selected as it encompasses several factors into a single index. The most deprived areas are likely to have high population density, minimal accessible greenspace and less mobility. Therefore, it is expected that moorland fringe close to deprived areas will see high footfall and elevated risk of ignition.

9.1.8. Lay-bys

It is known that stakeholders believe proximity to car parks influences wildfire distribution. However, in the previous report no distance decay relationship was found. Therefore, we felt the inclusion of lay-bys may help distinguish a relationship between car parking locations and wildfires. OS MasterMap cartographic text data was used to identify locations of lay-bys (point locations were extracted for every 'lay-by' label in the MasterMap dataset). A visual inspection of aerial imagery shows that not all lay-bys are identified using this methodology, but it was deemed the best approach given the constraints.

9.1.9. Car Parks

The same methodology for identifying lay-bys was used for car parks.

9.2. Annex 2: Detailed Methodologies

The MCE analysis was updated for the two highest performing models in McMorrow and Lindley (2006), models 1c and 2g. Whilst the weightings of each layer were maintained from McMorrow and Lindley (2006), it is important to note that these weights are likely not wholly appropriate and model results would be more robust if these weightings were revised.

Each habitat class was given an area weighted risk score. First, the habitat at each fire point was extracted. Second, the percentage of the study area covered by each class was calculated and used to predict the expected number of fires in each class. Residuals were calculated by subtracting the expected number of fires from the number observed. Scores were allocated to each class by awarding a score of 10 to largest positive residual, 1 to the largest negative residual and scaling the intervening values between 10 and 1.

The scored distance decay rasters for each predictor were updated by producing histograms as in McMorrow and Lindley (2006). These give the frequency of wildfires (during 2009-2018) within 'x' meter distance classes from each of the human factors included in the models. Distance class scores were guided by those given in McMorrow and Lindley (2006).

The histograms produced and the resulting scored distance classes are given for each dataset below.

9.2.1. Settlements

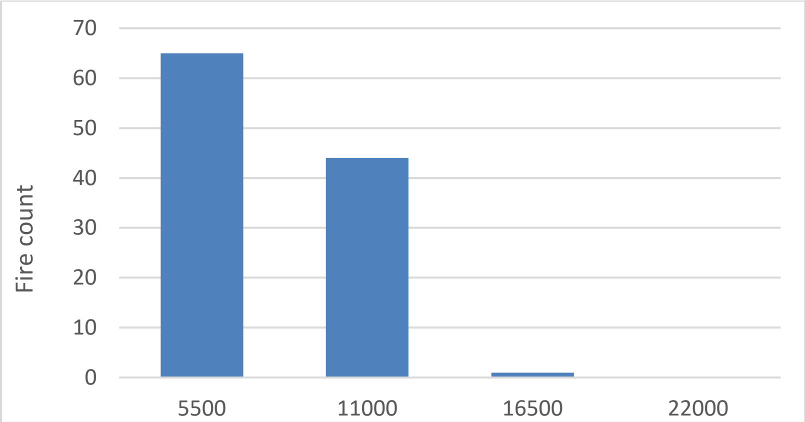

Figure 8: Frequency of wildfires within 5,500m distance classes from settlements. Bin labels on distance axis refer to right hand tick point.

The final scores assigned were:
0-5500m = 10
5500-11000m = 8
> 11000m = 0

9.2.2. Minor Roads

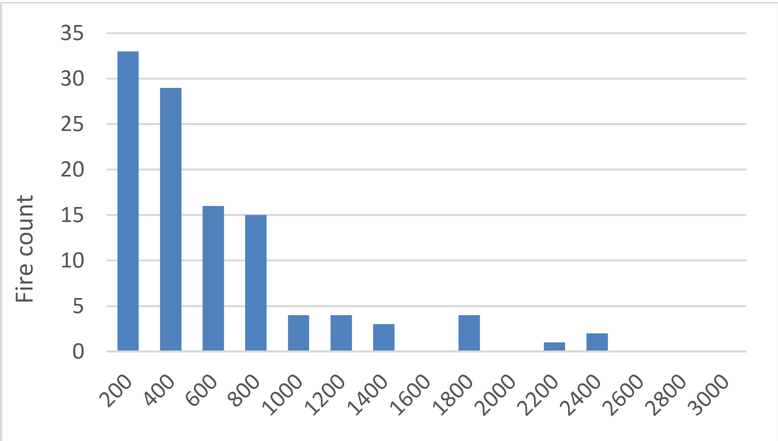

Figure 9: Frequency of wildfires within 200m distance classes from minor roads. Bin labels on distance axis refer to right hand tick point.

The final scores assigned were:
0-200m = 10
200-400m = 9
400-800m = 5
800- 2400 = 1
>2400m = 0

9.2.3. Pennine Way High Popularity

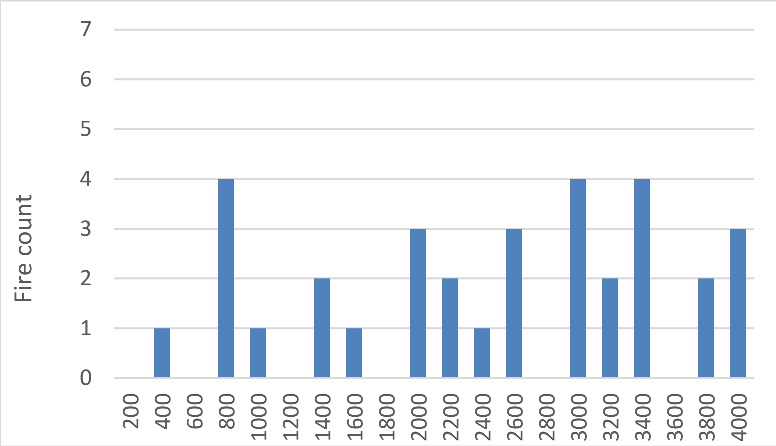

Figure 10: Frequency of wildfires within 200m distance classes from high popularity sections of the Pennine way. Bin labels on distance axis refer to right hand tick point.

The final scores assigned were:
All distances: 0
Note: All distances given 0 as there was no decrease in number of fires with distance from popular sections of the Pennine Way, in direct contrast to the 2006 report.

9.2.4. Public Rights of Way

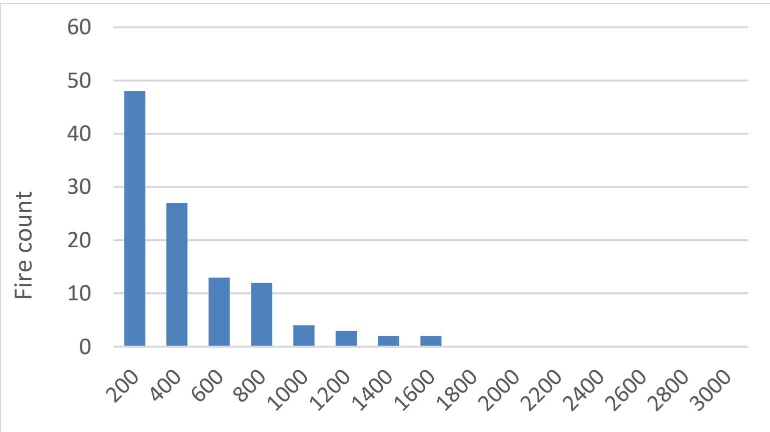

Figure 11: Frequency of wildfires within 200m distance classes from public rights of way. Bin labels on distance axis refer to right hand tick point.

The final scores assigned were:
0-200m = 10
200-400m = 6
400-800m = 3
800-1600m = 1
>1600m = 0

9.2.5. Waylines

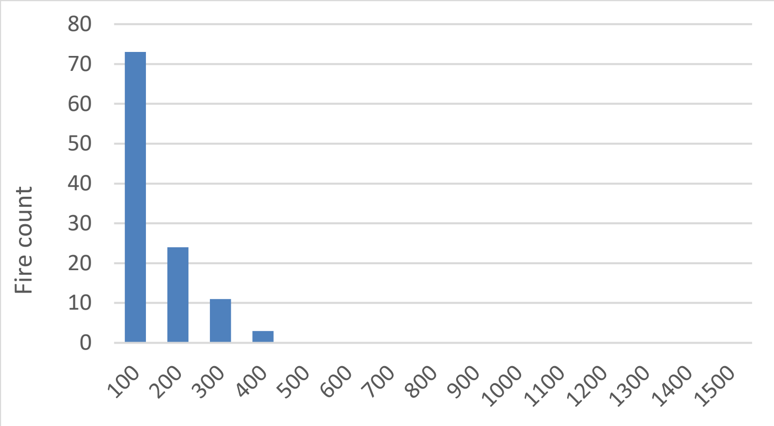

Figure 12: Frequency of wildfires within 100m distance classes from waylines. Bin labels on distance axis refer to right hand tick point.

The final scores assigned were:

0-100m = 10
100-200m = 3
200-300m = 2
300-400m = 1
>400m = 0

9.2.6. Land Cover

The CEH Land Cover Map habitat dataset was scored using the same procedure McMorrow and Lindley (2006) used to score the ESA habitat. An area weighted risk score was calculated for each of the 21 LCM classes. First, the habitat at each fire point was extracted, second the percentage of the study area covered by each class was calculated and used to predict the expected number of fires in each class. Scores were allocated to each class by awarding a score of 10 to largest positive residual, 1 to the largest negative residual and scaling the intervening values between 10 and 1 (Table 5).

Table 5: Scores attributed to LCM 2015 habitat classes

Habitat Class	Score
Heather Grassland	10
Heather	7
Broadleaved Woodland	6
Coniferous Woodland	6
Calcareous Grassland	6
Acid Grassland	4
Bog	1
Arable and Horticulture	0
Improved Grassland	0
Inland Rock	0
Freshwater	0

Suburban	0

The two models were run using the raster calculator in ArcGIS. The equations used to build the models were identical to those used in McMorrow and Lindley (2006), as the weightings remain the same. The equations used are below:

Model 1c
("LCM2015.tif" * 0.50) + ("Settlement.tif" * 0.1025) + ("Minor.tif" * 0.0225) + ("PWHiPop.tif" * 0.18) + ("PROW.tif" * 0.13125) + ("Wayline.tif" * 0.06375)

Model 2g
("LCM2015.tif" * 0.50) + ("Settlement.tif" * 0.3075) + ("Minor.tif"* 0.0675) + ("PWHiPop.tif" * 0.06) + ("PROW.tif" * 0.04375) + ("Wayline.tif" * 0.02125)

9.2.7. Logistic Regression

The Bayesian Information Criterion (BIC) was used to select variables using backward stepwise selection. BIC selects a more parsimonious model, reducing the likelihood of overfitting. Overfitting occurs when the model begins to fit the random error of the sample, rather than the relationship between variables in the real world. It is particularly a risk when sample size is small, hence the decision to be conservative and leave lay-bys out of the final model. To build the 10 models from which model variables were selected, 10 random samples of 222 non ignition points were selected. The same 111 ignition points (all known fires in the database during the period) were used in each of the 10 models.

9.3. Annex 3: Logistic regression performance metrics

As an example, the performance metrics of the model derived from the 10th sample are provided below (10 samples were taken to derive average coefficients for the final model).

Table 6: Coefficient estimates, standard error and significance of variables in model built using sample no. 10

	Coefficient	Std. error	Z value	Pr(>|z|)	Significance
Intercept	1.489	0.427	3.487	0.000	0.00
Dist_IMD_1to3	-0.0002154	0.0000668	-3.221	0.001	0.01
Dist_Minor	-0.0009865	0.0003105	-3.177	0.001	0.01
Dist_Wayline	-0.003787	0.001760	-2.151	0.031	0.05

Table 7: Statistical tests from model built using sample no. 10. A 0.5 cut off was used to distinguish between non-ignition/ignition)

Test	Score
AUC	0.823
Accuracy (Acc)	0.752
No Information Rate (NIR)	0.624

P-value (Acc > NIR)	0.000
Kappa	0.418
Mcnemar's test P value	0.000
Sensitivity (true positive rate)	0.940
Specificity (true negative rate)	0.440

Table 8: Confusion matrix from model built using sample no. 10. A 0.5 cut off was used to distinguish between non-ignition/ignition)

	Predicted non-ignition	Predicted ignition
Actual non-ignition	78 (True Negative)	28 (False Positive)
Actual ignition	5 (False Negative)	22 (True Positive)

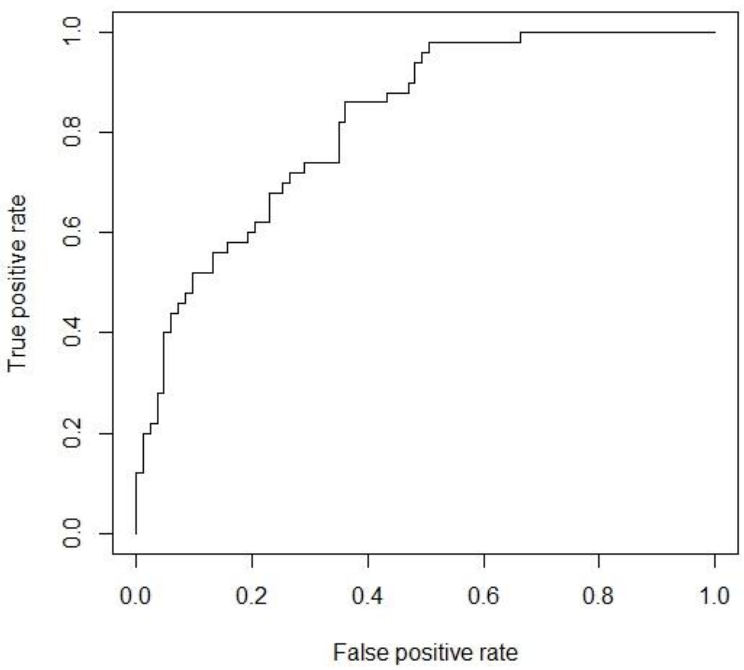

Figure 13: ROC plot from model built using sample no. 10. Note true positive rate is equivalent to actual fires, and false positive rate is equivalent to false alarms.

9.4. Annex 4: R code used to build and assess logistic regression models

```
# install and load required packages

install.packages("corrplot")
install.packages("caret")
install.packages("regclass")
install.packages("raster")
```

```
install.packages("ROCR")
install.packages("pROC")
install.packages("rgdal")
install.packages("e1071")

library(raster)
library(dplyr)
library(ROCR)
library(pROC)
library(corrplot)
library(caret)
library(regclass)
library(rgdal)
library(e1071)

############################

# set working directory
setwd("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Rebuilding_Logistic")

# create file list to create raster stack
fs <- list.files(path =
"U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Rebuilding_Logistic\\Raster_inputs", pattern
= "tif$", full.names = TRUE)

# create raster stack using file list
rasstack <- stack(fs)

###########################

# Create non ignition sample points
# load non ignition area
# Non ignition area was created in ArcMap. It is the study area minus ignition loacation plus
a 200m buffer.
non_ignit_area <-
readOGR(dsn="U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition
_test", layer="Non_ignit_area_200m")

# set seed
# this was changed for each sample run
set.seed(1)

# create random sample points in non ignition area
NonIgnitSample <- spsample(non_ignit_area,n=222,"random")

# write random sample coordinates to csv
```

```
write.csv(NonIgnitSample@coords, file =
"U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test\\nonigni
tpoints1.csv", row.names = FALSE)

#####################

# load ignition and non ignition points
# Ignition points derive from the MFFP wildfire database
Ignition <-
read.csv("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Rebuilding_Logistic\\Sample_points
\\Ignition_points_09to18.csv")
Non_Ignition <-
read.csv("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test
\\nonignitpoints1.csv")

# set csv files to spatial points data frame
coordinates(Ignition) <- ~ Point_X + Point_Y
coordinates(Non_Ignition) <- ~ x + y
Non_Ignition <- as(Non_Ignition,"SpatialPointsDataFrame")

# extract raster values from sample points
Ignition_values <- extract(rasstack,Ignition)
Non_Ignition_values <- extract(rasstack,Non_Ignition)

# Combine raster values with points
Ignition_Comb <- cbind(Ignition,Ignition_values)
Non_Ignition_Comb <- cbind(Non_Ignition@coords,Non_Ignition_values)

# write combined files to csv for future reference
write.csv(Ignition_Comb, file =
"U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test\\Ignition
_rast_values.csv", row.names = FALSE)
write.csv(Non_Ignition_Comb, file =
"U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test\\Non_Ig
nition_rast_values_200m_v1.csv", row.names = FALSE)

#########################

# Add ignition and non ignition raster data to environment ready for combination
Ignition_csv <-
read.csv("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test
\\Ignition_rast_values.csv")
Non_Ignition_csv <-
read.csv("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test
\\Non_Ignition_rast_values_200m_v1.csv")

# Add column to give ID for whether sample point is an ignition source or not
```

```
Ignition_csv['Ignition'] = 1
Non_Ignition_csv['Ignition'] = 0

# Combine the two datasets to give the final data to build the model
Build_data <- rbind(Ignition_csv, Non_Ignition_csv)

# write the build data to csv for future reference
write.csv(Build_data, file =
"U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test\\Model_
build_data_09-18_200m_v7.csv", row.names = FALSE)

# if already built input data can load from file
# Build_data <-
read.csv("U:\\GIS\\Projects\\Fire_Risk_Map\\Analysis\\Testing_Logistic\\Non_Ignition_test
\\Model_build_data_09-18_200m_v1.csv")

# view correlation matrix between all predictor variables
correls <- cor(Build_data[,2:10])
corrplot(correls, method = "circle")

# extract training and testing data
set.seed(2018)
n <- nrow(Build_data)
shuffled_df <- Build_data[sample(n), ]
train_indices <- 1:round(0.6 * n)
train <- shuffled_df[train_indices, ]
test_indices <- (round(0.6 * n) + 1 ):n
test <- shuffled_df[test_indices, ]

######################################

# build model using all variables ready for the backwards selection
allvar <- glm(Ignition ~ dist_CarPark + dist_IMD1to3 + dist_LayBy + dist_Major + dist_Minor
+ dist_PROW + dist_PWHiPop + dist_Wayline + LCM2015_clipped, data = train, family =
binomial())

# summary stats of the model
summary(allvar)

# Run the backwards stepwise regression
# using k=log(n) which uses the BIC function for selecting the model.
# using the BIC functiuon as it seems to be quite stringent on what it includes
backwards <- step(allvar, k=log(n))
summary(backwards)

# calculate and plot ROC AUC on test data
p <- predict(backwardsselect, newdata = test, type = "response")
```

```
pr <- prediction(p, test$Ignition)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

# build confusion matrix
confusionMatrix(table(predict(backwardsselect, newdata = test, type = "response") >= 0.5,
        test$Ignition == 1))
```