Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets

Article

Accepted Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

1 **Mapping and validating predictions of bacterial biodiversity using European and**

2 **national scale datasets**

3

4 Robert I. Griffiths[1], Bruce C. Thomson[1], Pierre Plassart[2], Hyun S Gweon[1], Dorothy Stone[3],

5 Rachael E. Creamer[3], Philippe Lemanceau[4], Mark J Bailey[1]

6

7 [1]Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford,

8 Wallingford, UK

9 [2]INRA, UMR1347 Agroécologie, GenoSol Platform, Dijon, France

10 [3]Teagasc, Johnstown Castle Research Centre, Co. Wexford, Ireland

11 [4]INRA, UMR1347 Agroécologie, Dijon, France

12

13 Corresponding author: R.I.G (email: rig@ceh.ac.uk)

14

15

16

17

18

19

20

21

22

Abstract

Recent research has highlighted strong correlations between soil edaphic parameters and bacterial biodiversity. Here we seek to explore these relationships across the European Union member states with respect to mapping bacterial biodiversity at the continental scale. As part of the EU FP7 EcoFINDERs project, bacterial communities from 76 soil samples taken across Europe were assessed from eleven countries encompassing Arctic to Southern Mediterranean climes, representing a diverse range of soil types and land uses (grassland, forest and arable land). We found predictable relationships between community biodiversity (ordination site scores) and land use factors as well as soil properties such as pH. Based on the modelled relationship between soil pH and bacterial biodiversity found for the surveyed soils, we were able to predict biodiversity in ~1000 soils for which soil pH data had been collected as part of national scale monitoring. We then performed interpolative mapping utilising existing EU wide soil pH data to present the first map of bacterial biodiversity across the EU member states. The predictive accuracy of the map was assessed again using the national scale data, but this time contrasting the EU wide *spatial* predictions with point data on bacterial communities. Generally the maps were useful at predicting broad extremes of biodiversity reflective of low or high pH soils, though predictive accuracy was limited for Britain particularly for organic/acidic soil communities. Spatial accuracy could however be increased by utilising published maps of soil pH calculated using geostatistical approaches at both global and national scales. These findings will contribute to wider efforts to predict and understand the spatial distribution of soil biodiversity at global scales. Further work should focus on enhancing the predictive power of such maps, by harmonising global datasets on soil conditioning parameters, soil properties and biodiversity; and the continued efforts to advance the geostatistical modelling of specific components of soil biodiversity at local to global scales.

1. Introduction

49   Soil bacteria contribute the largest proportion of the soil genetic resource (Urich et al

50   2008; Fierer et al, 2012), reflecting their ubiquity and high abundance across all soil systems.

51   Given bacterial importance in the regulation of soil ecosystem services (Comerford et al,

52   2013), increased understanding of the environmental controls of bacterial biodiversity is

53   required from both scientific and policy perspectives in order to predict biodiversity change,

54   and determine functional consequences of change due to future climatic or land use

55   pressures. Attempts to characterise the bacterial communities in soils and understand

56   ecological drivers have previously been hampered by methodological difficulties in assessing

57   taxonomic diversity due to the limited culturability of many bacterial taxa coupled with vast

58   taxonomic diversity (e.g Janssen et al, 2002). These problems have to some extent been

59   overcome through the development of molecular technologies to assess the diversity of

60   taxonomic marker genes (particularly the 16S rRNA gene) PCR amplified from extracted soil

61   DNA (Hirsch et al, 2010).

62   The application of molecular methods to wide ranging globally dispersed soil samples

63   has revealed that soil bacterial communities are broadly structured along gradients of soil

64   properties, with strong correlations between measures of bacterial biodiversity and key soil

65   variables such as soil pH and organic matter, which are co-related with broader environmental

66   parameters such as land use, climate, and parent material (Fierer et al, 2006; Lauber et al,

67   2009; Griffiths et al 2011). Therefore, whilst the causal mechanisms underlying these

68   relationships are complex it is apparent that the same pedogenic factors which determine the

69   nature of soils (e.g Jenny, 1941) also determine the taxonomic characteristics and structure

70   of the soil bacterial community. This new knowledge permits spatial forecasting of bacterial

71   biodiversity at a range of scales and under change scenarios; which together with parallel

72   developments in understanding microbial biodiversity-function relationships, may allow for

73   enhanced prediction of soil processes under future environmental change.

74   Molecular surveys permit the production of range maps of soil bacterial distributions

75   at various spatial scales. Spatial distribution maps provide a visual representation of the

76   forces shaping populations or communities and therefore provide the foundation for macro

77   ecological understanding (Elton, 1927). Maps can also guide policy decisions with respect to

78   land management, and can be useful visual resources guiding scientific experimentation and

79   enquiry. Importantly, more recently rasterised maps provide georeferenced data which can

80   feed wider ecological, climatic or biogeochemical models. Already there has been several

81   attempts to map soil microbial properties at national and regional scales, using molecular

82   methodologies applied to nationwide soil monitoring schemes (Bru et al, 2011; Griffiths et al

83   2011, Dequidet et al 2009; Dequidet et al 2011). These studies mapped point sampled

84   microbial data using *interpolative* methods (e.g inverse distance weighting, kriging; see

85   Bivand et al; 2008) to fit surfaces predicting the microbial properties at unsampled locations

86   by weighted averages of surrounding measured values. These methods are useful to show

87   large differences in microbial properties over large areas but local accuracy is limited by the

88   spatial scale of sampling.

89   More advanced geostatistical approaches can be used to predict a variable of interest

90   at unsampled locations based on known relationships between the dependant variable and

91   other predictor variables (e.g climate, soil type, land cover). Such approaches are commonly

92   used in wider ecology (sometimes termed environmental-, ecological-, or species- distribution

93   modelling: Elith et al, 2006), and can be used to predict either species or communities at

94   unsampled locations (Chapman and Purse, 2011). These environmental correlational

95   approaches have so far been used to predict historical change in soil bacterial biodiversity

96   due to land use at regional scales (Fierer et al, 2013); and also to improve on the interpolated

97   maps of bacterial biodiversity across Great Britain (Griffiths et al, 2011) by modelling the

98   observed relationships between bacterial communities and environmental variables, and then

99   forecasting communities in unsampled locations using remote sensed land cover information

100   and parent material maps (Henrys et al, 2015). This paper aside there are few studies which

101   have examined in detail the predictive performance of such maps compared to simple

102   interpolation. More widely, large scale spatial predictions of soil parameters are increasingly

103   being disseminated through downloadable map resources (e.g soilgrids.org, ukso.org), and

104   there is now a need to identify specific predictive limitations in order to further improve

105   accuracy (Hengl et al, 2014).

106   Here as part of this special issue reporting results from the EU FP7 EcoFINDERs

107   project coordinated soil sampling campaign, we seek to assess the bacterial communities in

108   76 soils sampled across Europe in order to produce a soil bacterial map at the European

109   scale, which can be validated against national scale datasets. We predict that soil pH will be

110   the strongest correlate with measures of community biodiversity, which will then allow us to

111   predict and spatially interpolate communities based on publicly available European scale

112   point data on soil pH (from the LUCAS survey: Tóth et al, 2013). The predictive accuracy of

113   this map will be assessed by comparing predictions with observed point data on bacterial

114   communities collected with similar methods from over 1000 soils across Great Britain (Griffiths

115   et al, 2011). We will also explore whether the predictions from this simple interpolated map

116   can be improved upon, by spatially predicting communities based on existing soil pH maps

117   produced using more advanced environmental correlation approaches (from soilgrids.org and

118   ukso.org).

119

120   **2. Materials and Methods**

121   **2.1. Sampling**

122   Bacterial communities were examined in soils sampled across the EU member states

123   as part of the EcoFINDERs project "transect" sampling campaign, full details of which are

124   prodived elsewhere in this issue (Stone et al, 2015). Briefly, a range of sites spanning a

125   gradient of soil properties (principally pH, organic matter and texture), climatic zones, and

126   land uses (grassland, arable, forest) were targeted for sampling following examination of EU

127   wide datasets (see supplementary material for site locations, S1). Samples were collected at

128   the end of summer 2012 according to standardised protocols to 5cm depth, and sent to a

129   central processing lab for homogenisation and distributing to various partner labs for further

130   analyses. In total eighty-two soils from 11 countries encompassing Arctic to Southern

131   Mediterranean climes of which 76 are assessed in this study. Soil chemical determinations

132   were also conducted by a single laboratory to provide measures of volumetric moisture

133   content, pH (in water), texture, and total/organic carbon (C) and nitrogen (N) contents.

134   **2.2. DNA extraction and community analyses**

135   Total genomic DNA was extracted from all soil samples using a previously described

136   DNA extraction procedure (Plassart et al., 2012). Briefly, 1g of soil was mixed at 70°C with a

137   extraction buffer containing 100 mM Tris-HCl (pH 8), 100 mM EDTA (pH8), 100 mM NaCl,

138   2% (w/v) polyvinylpyrrolidone (40 g mol-1) and 2% (w/v) sodium dodecyl sulphate. Proteins

139   were precipitated from the supernatant with 1/10 volume of 3 M sodium acetate, before

140   nucleic acid precipitation with isopropanol. DNA was further purified through

141   polyvinylpolypyrrolidone (PVPP) Microbiospin minicolumns (BIORAD, Marnes-la-Coquette,

142   France) and finally using the Geneclean Turbo kit (MP-Biomedicals, NY, USA).

143   Bacterial communities were examined using TRFLP as described by Griffiths et al

144   (2011) using the forward primer 63F (5′-CAGGCCTAACACATGCAAGTC-3′) labelled at the

145   5′ end with D4 blue fluorescent dye and reverse primer 530R (5′-GTA TTA CCGCGG CTG

146   CTG-3′). Amplifications were performed in 50 µl reactions under the following conditions:

147   94^0C for 90 s, followed by 35 cycles of 94^0C for 45 s, 55^0C for 1 min and 72^0C for 3 min,

148   followed by a final extension of 72^0C for 10 min. Amplicons were then purified using the ZR-

149   96 DNA clean-up kit (Zymo research, Freiburg, Germany), prior to enzymatic digestion.

150   Purified bacterial DNA was digested with MspI restriction enzyme (New England Biolabs Inc.,

151   Ipswich, MA, USA) at 37^0C for 3 h. Fragment analysis was performed with a Beckman Coulter

152   CEQ 2000XL capillary sequencer (Beckman Coulter Corporation, California, USA). Peak

153   height data were analysed using GeneMarker software (Softgenetics, LLC, PA, USA).

154   Relative abundances were calculated as the ratio between the fluorescence of each terminal

155   restriction fragment (T-RF) and the total integrated fluorescence of all T-RFs.

156   **2.3. Statistical Analyses**

157   A site by taxon (TRF) relative abundance table derived from the TRFLP analyses was

158   used to explore community relationships with environmental variables, and calculate

159   community scores (ordination site scores and diversity estimates) using standard routines in

160   the vegan library within the R package (R Core Development Team, 2005). Geostatistical

161   calculations, manipulations and plots were also performed within R using the maptools, gstat,

162   raster, and RColorBrewer libraries. Specifically, to produce the bacterial map we used the

163   inverse distance distance weighted (IDW) interpolation method, on account of it's simplicity

164   and widespread application (Lam, 1983). The IDW method predicts a value at an *unsampled*

165   location based on the weighted average of values at *sampled* point locations, with weights

166   decreasing linearly with distance from that location. We used the idw function of the R library

167   gstat to perform the interpolation, using leave one out cross validation to establish the

168   optimum power parameter value (determining how much the weightings decrease with

169   distance) and evaluate the overall performance of the interpolation with respect to predictive

170   power. For both the IDW interpolative mapping, and the prediction contrasts with observed

171   data from a national scale dataset, predictive power was evaluated by assessing the

172   coefficient of determination (R^2) and root mean square error (RMSE) between observed and

173   predicted values.

174

175

176   **3. Results and Discussion**

177   **3.1 Continental scale patterns of soil microbial communities**

178   NMDS ordinations revealed distinct clustering of sampled communities according to

179   land use type (Figure 1). This was further confirmed following multivariate permutation tests

180   using the anosim statistic (R = 0.28, P = 0.0001). Pairwise comparisons further revealed that

181   bacterial communities in forest soils were most distinct from to arable and grass communities

182   (R=0.54, and R=0.41 respectively, p<0.0001) with the largest differences in community

183   structure consistently observed between forest and arable soils. Arable and grass

184   communities were more similar, yet significant differences were still apparent despite the wide

185   dispersion at the continental scale of sampling units (R=0.08, p<0.05). Bacterial communities

186   were found to differ between countries (R = 0.13, P < 0.01). However, this effect could be

187   predominantly attributed to the Swedish soil communities which were all sampled from forest

188   sites and formed a distinct outgroup in the ordination (Figure 1b). When Swedish samples

189   were excluded country of origin had no significant effect (bacteria R = 0.001, P = 0.46).

190   Fitting of environmental variables to the ordination scores also confirmed that

191   microbial communities sampled across Europe were strongly correlated with environmental

192   gradients. The dominant five environmental conditions most strongly associated with

193   microbial community structure differences are presented in Table 1. Generally, bacterial

194   community differences were highly correlated with change in soil chemistry and nutrient

195   status, with soil pH showing the strongest relationship, confirming that across large spatial

196   scales the structuring of soil bacterial communities is largely predictable by common soil

197   physicochemical parameters.

198   These findings further highlight the difficulty in separating direct effects of land use on soil

199   biota versus indirect effects, mediated by changes in soil abiotic properties. It is becoming

200   increasingly apparent that none of these parameters are independent. Human land use is

201   generally influenced by the local pedo-climatic context which determines the economic

202   suitability of different land management options. The baseline pedo-climatic state will naturally

203   create topsoils of distinct properties, which can be further modified by land use depending on

204   the specific intervention. Different land uses are therefore often accompanied by distinct

205   abiotic soil properties and consequently bacterial biodiversity, given the strong relationships

206   between edaphic properties and soil bacterial communities. For instance, Scandinavian

207   regions are characterised by cold conditions and acidic soils giving rise to more forest and

208   less arable suitability. This in itself does not mean that soil bacterial communities are

209   inherently geographically structured, nor that they are fundamentally driven by the land use

210   of forestry, but is more a reflection of the natural pedo-climatic state which determines both

211   the human land use and the soil biotic and abiotic properties. With respect to contrasts

212   between arable and grassland habitats; whilst arable soils are generally defined by a relatively

213   narrower set of soil properties (e.g high pH and low organic matter) it is possible for grasslands

214   to possess similar properties, particularly if the grassland is part of a arable rotation. Such

215   historical data is not available in this study and such specific contrasts are better addressed

216   in locally focused long term experimental contrasts.

217   **3.2 Predictability and mapping of soil bacterial communities**

218   The site scores for bacterial communities were clearly strongly aligned along the first

219   axis of the NMDS ordination which corresponds with a gradient of soil pH. This afforded the

220   opportunity to extrapolate and predict communities over larger spatial scales using wider soil

221   pH datasets. Such datasets are available across 23 EU member states from the LUCAS

222   topsoil survey (Toth et al, 2013), which provides data on the percentage of coarse fragments,

223   particle size distribution, pH, organic carbon, carbonate content, phosphorous content, total

224   nitrogen content, extractable potassium content, cation exchange capacity and multispectral

225   properties from approximately 20000 soils. We therefore sought to model the relationships

226   between soil pH and bacterial communities from the present survey, and then predict

227   community NMDS scores for the 20000 data points across the EU of soil pH to enable the

228   production of EU wide maps of predicted soil biodiversity using simple interpolative

229   approaches.

230   The first step was to reliably model relationships between soil pH and the bacterial

231   NMDS scores. Figure 2 shows the relationship observed between soil pH and the first axis

232   bacterial community NMDS scores for the 76 soils assessed in this study. The relationship

233   was visually assessed to be curvilinear and could be modelled with a simple second-degree

234   polynomial (R^2 = 0.93) of the equation:

235   bacterial $NMDS^{axis1}$= -3.748 + 0.9188pH – 0.04954pH^2

236   To test the predictive power of the regression equation, we predicted the bacterial

237   community NMDS axis 1 scores from a nationwide survey of over 1000 point measurements

238   conducted across Great Britain using only the measured pH values as predictors

239   (countrysidesurvey.org.uk/). Uniquely, this dataset also comprises bacterial TRFLP profiles

240   (Griffiths et al 2011) therefore allowing us to independently test the predictive power of the

241   regression equation on a different dataset. Despite several differences in methodologies (soil

242   sampling depth, DNA extraction, taxonomic binning) the modelled relationships between soil

243   pH and bacterial community ordination scores from less than 100 soils across Europe

244   provided a reasonable prediction of the bacterial scores in over 1000 soils across Britain

245   (Figure 3). It is noteworthy that the ordination axis scores are themselves arbitrary, and only

246   denote the (dis)similarity between samples analysed at any one time. This fact makes the

247   strong correlations between the EU wide and national scale datasets all the remarkable, and

248   is perhaps testament to the strength and global ubiquity of the relationships between soil pH

249   and bacterial communities, provided a sufficient range of soils are sampled.

250   To map bacterial communities across the EU member states we then predicted the

251   NMDS axis 1 ordination scores for the ~20000 soils sampled in the LUCAS survey based on

252   pH measurements and the equation outlined above. The LUCAS datasets were downloaded

253   subject to agreements from the JRC European Soil Portal (http://eusoils.jrc.ec.europa.eu/)

254   and a map showing the sampled locations is provided in the supplementary material (S2).

255   Predicted community scores were then mapped using inverse distance weighting

256   interpolation. We firstly compared interpolative performance using different integer powers

257   parameters (1-5), and the accuracy of predictions assessed by comparing the deviation from

258   the measured data using a leave one out cross-validation procedure. The best performing

259   interpolated map is shown in Figure 4. This was made using an IDW power parameter of 2

260   which yielded the lowest root mean square error (RMSE) for predicting bacterial NMDS1

261   scores (0.28), together with the highest precision with respect to the relationship between

262   observed and predicted values (R^2=0.58).

263   **3.3 Features of the map**

264   The map reveals the broad types of bacterial communities found across Europe based

265   on the strong relationships between soil bacterial biodiversity and soil pH. The low (negative)

266   NMDS axis 1 scores reflect communities found in areas such as Scandinavia where acidic

267   and organic rich soils develop due to climatic factors; whereas high values indicate

268   communities found in more productive Southern circum-neutral pH soils, typically with lower

269   organic matter. Areas of contrasting local variability can also be seen in certain regions, where

270   geological factors such as differences in underlying parent material or topography cause local

271   change in communities.

272   To taxonomically interpret the features of the map we must firstly consider the

273   "meaning" of the first axis ordination scores. The axis 1 ordination scores summarise

274   differences in the broad taxonomic composition and relative abundance of taxa between

275   samples. Additionally, in this study, the scores correlated positively with indices of diversity

276   (Figure 5), with lower scores reflecting low taxonomic diversity. The specific change in

277   abundance of different bacterial taxa across soil pH/diversity gradients has been well studied

278   using sequencing (e.g see Lauber et al 2009; Rousk et al 2010; Griffiths et al, 2011) and can

279   also be inferred to some extent from TRFLP analyses (some illustrative responses of

280   dominant TRFLP peaks are shown the supplementary material, S3). To summarise these

281   responses briefly, acidic and organic rich soils are notably dominated by distinct lineages of

282   acidophilic acidobacteria, as well as alphaproteobacterial taxa. As pH increases over soil

283   environmental gradients, the alphaproteobacteria become dominant, followed by other broad

284   taxonomic groups such as other proteobacteria and the actinobacteria as pH nears neutrality.

285   Neutral soils typically comprise a wider variety of different bacterial taxonomic groups

286   of higher evenness (Griffiths et al, 2011); a phenomena which is at odds with the notion that

287   agricultural soils (often neutral pH) are depauperate with respect to biodiversity (Spurgeon et

288   al, 2013). Potential explanations for higher soil bacterial biodiversity in agricultural soils could

289   be: i) more bacterial taxa exist at neutral pH, due to the requirements of intracellular pH

290   homeostasis (Booth, 1985); ii) soil physical properties in mineral agricultural soils provide

291   more microhabitats promoting evenness (e.g spatial isolation theories, Zhou et al, 2002); and

292   (iii) mineral agricultural soils have less active populations meaning a high diversity of

293   sensescent cells, or even extracellular DNA are being detected. We note also that several

294   studies have reported increased dominance of certain lineages, despite higher phylogenetic

295   diversity, in neutral soils resulting in declines in indices of diversity at higher pH (e.g Fierer et

296   al, 2006). This is apparent to some extent in certain arable and grassland soils in Figure 5

297   which appear to have a marked dominance of alphaproteobacterial TRF peaks, though the

298   underlying causes of this have yet to be fully elucidated. Given the importance of these soils

299   for agricultural production together with recent concerns over soil and food security, the

300   specific controls of neutral-soil taxon abundances, and functional consequences of alterations

301   in abundance, represents a key current knowledge gap.

302

303   **3.4 Map validation and contrasts with other mapping approaches**

304   In order to assess the accuracy of the EU bacterial map wide we contrasted the spatial

305   predictions with observed national scale data from the British Survey. The interpolated map

306   was firstly converted to a raster, and then the predicted NMDS1 scores extracted using the

307   sample locations of the British dataset, prior to correlation with the observed scores (figure

308   6a). Despite a lack of a good linear fit and a tendency for the spatial predictions to cluster

309   near the overall mean, the interpolation performed reasonably well at predicting the NMDS

310   community scores across Britain (RMSE=0.41, R^2 = 0.29). Despite a lack of strong correlative

311   relationships for lower pH communities, there was evidence that higher scoring (higher pH)

312   community scores could be predicted to some extent. This map of bacterial biodiversity

313   therefore gives a very broad overview of the extreme types of communities likely to be found

314   in different geographic locations across Europe, but for Britain it is of limited use in spatially

315   predicting more subtle differences in communities. Its predictive power is limited by its reliance

316   on the locations of the sampled LUCAS topsoil data points, the design of which has an

317   inherent bias towards agricultural lands (Toth et al 2013). Few samples were taken from large

318   areas of Scottish uplands in the LUCAS survey which may be explain the poor relationships

319   between predicted and observed community scores across Britain. The lack of comparable

320   national scale "test" datasets comprising bacterial data, means we are unable to assess the

321   predictive accuracy of the map for other countries.

322   To assess whether the predictive accuracy could be enhanced by drawing on more

323   advanced geostatistical predictions of soil pH, we next applied the pH-biodiversity transfer

324   function to two existing soil pH maps: a recently published predictive map at the global scale

325   (SoilGrids: soilgrids.org, Hengl et al 2014) and freely available maps of soil pH at the national

326   scale from Britain (Countryside Survey data from ukso.org). Both these maps were

327   constructed using geostatistical models applied to surveyed pH data to predict unknown

328   values using wider landscape level datasets, using information such as land cover, parent

329   material, climate etc. Maps were downloaded and rasterized where necessary, prior to

330   extracting of pH values based on the GB survey coordinates.

331   Using the detailed 1km resolution SoilGrids soil pH map offered some small

332   improvements in predicting the national scale bacterial data (RMSE=0.39, R^2 = 0.36)

333   particularly with respect to the acidic habitat scores (Figure 6b). However the predictions

334   again were focused around the mean, and extreme scores were not particularly well

335   estimated. It was notable that in inspecting the range of soil pH values predicted across the

336   UK and comparing with known UK level data that the extreme values were particularly

337   underestimated in the SoilGrids predictions (e.g predictions of pH 4 or pH 8 soils were over

338   or underestimated respectively). Possible explanations could be related to i) difference in pH

339   determination between the Countryside survey and EU wide LUCAS datasets; ii)

340   undersampling of certain habitats at the EU scale; and iii) geo-statistical artefacts. It is

341   impossible to entirely discount (i) as comparable samples are not available from both surveys,

342   but a cursory inspection of the range of pH values for both datasets indicated there were no

343   systemic differences in the range of pH measurements. With respect to ii) as already

344   discussed, part of the reason for the poor fit on the negative side of the interpolated map is

345   the lack of upland areas sampled in the Lucas survey – meaning the predicted pH for under

346   sampled upland areas such as in Scotland would be higher than in reality. Whilst the SoilGrids

347   predictions utilise global soil pH data in the model, the LUCAS dataset is also a large

348   contributory dataset which could have a significant influence on the predictions. Finally with

349   respect to geostatistical artefacts (iii), the predictive maps of pH provide a mean prediction

350   based on a global model, which will always under or overestimate the higher or lower

351   extremities of the pH range respectively. This could explain why the full range of NMDS1

352   scores was not adequately reflected in the model predictions.

353   The 1km resolution national scale pH map performed considerably better in predicting

354   the range of community scores across the Britain (Figure 6c), with better coverage of the

355   extreme ends of the scale, and a much better fit overall (RMSE=0.31, R^2=0.60). However the

356   entire gradient of scores was not well reflected, since here the predictive map was calculated

357   based on modelled relationships between soil pH and categorical variables denoting land

358   cover, along with continuous variables related to parent material. Therefore only a limited

359   number of predicted pH categories are available in this map, constrained by the number of

360 land cover classes. Again we observed larger predictive error for lower pH soil communities,

361 which could relate to potentially weaker relationships between pH and available land cover

362 classes in these habitats, or less influence of parent material in these generally more organic

363 soils. Another possible reason for the larger relative error for lower scoring communities form

364 acidic habitats could relate to landscape patchiness and the resolution of the maps. For

365 instance, intensively managed parcels of land in the Britain are likely to comprise areas of

366 greater than $1km^2$ which is the scale of the UKSO pH map. Human intensified landscapes will

367 typically be more homogenous and of approximately neutral soil pH to favour plant production

368 (either "naturally" or agriculturally driven). This enhances the probability that a mean value

369 per km^2 will reflect a pH measurement at any given point within that square. Conversely,

370 marginal $1km^2$ land patches unfavourable for intensive agriculture will have a greater variety

371 of habitats and so the predictive accuracy with respect to point measures is likely to be

372 reduced.

373

374 **Conclusions**

375   This study characterised bacterial biodiversity and explored environmental correlates

376 in a range of soils sampled across continental Europe. In agreement with previous global

377 studies land use, climate and soil abiotic properties were strongly associated with changes in

378 bacterial communities, with soil pH being the best single correlate. Ultimately these findings

379 point to the general conclusion that broad characteristics of soil bacterial communities can be

380 considered as a dependent soil state variable related to other soil properties (and to some

381 extent human land use); which are ultimately controlled by the independent soil forming

382 factors of climate, relief, parent material, and time (Jenny, 1941). These relationships

383 therefore allow the global prediction of soil bacterial community features over large scales,

384 and we present the first attempt to map bacterial communities across Europe along with a

385 detailed evaluation of the predictions against observed data from national scale surveys of

386   one of the EU member states. The map performed adequately in predicting community

387   characteristics (ordination axis scores) albeit for opposing ends of the soil biotic/abiotic

388   gradient, and we further demonstrate how the map can be improved by making use of

389   available predictive maps of soil pH, previously calculated using correlations with

390   georeferenced data on wider soil forming factors. In doing so we highlight the current

391   limitations in soil property maps at different geographic scales, with national scale predictions

392   outperforming global scale maps.

393   To avoid misinterpretation there are some notable caveats which we must stress with

394   respect to the findings of this study and other large surveys of bacterial taxa utilising 16S

395   rRNA amplicon approaches. Firstly when conducting such large scale studies, one

396   necessarily focusses on broad patterns and, particularly in this study using a community

397   profiling technique, broad taxonomic resolution. We therefore do not propose that the map in

398   any way represents similarities between soils in terms of clonal or even species level

399   composition, which may be more governed by local ecological or evolutionary processes (Cho

400   and Tiedje, 2000). Additionally, it is be stressed that soil pH is not the sole driver of differences

401   in bacterial communities, nor should any form of causation be inferred. For instance the role

402   of plant inputs can also affect the relative abundances of taxa over relatively short timescales,

403   with potentially important functional consequences for processes such as carbon cycling

404   (Thomson et al, 2013).

405   Ultimately these taxonomic limitations will be overcome with wider global adoption of

406   sequencing approaches in soil monitoring networks which will likely enable environmentally

407   driven predictive models of individual taxon abundances (Fierer et al, 2013), rather than using

408   multivariate community estimates and relationships with well characterised soil biotic

409   variables. Whilst field-scale resolution was not the purpose of this mapping exercise, we feel

410   this should be a future ambition of global efforts to characterise soil biodiversity. Our study

411   highlights the benefits of using advances geostatistical approaches for soil biodiversity

412 mapping using high resolution remote sensed data. It is possible that similar approaches can

413 be applied to other elements of soil biodiversity, including eukaryotes, given enhanced

414 understanding of controlling environmental parameters. Such knowledge can be gained both

415 by efforts to harmonise existing soil biodiversity datasets, but also by increased eukaryotic

416 sampling in national surveys - a realistic possibility now with the availability of rapid molecular

417 tools for eukaryotes (e.g Ramirez et al, 2014). Finally to conclude, our map identifies that

418 predictions are only as good as the surveyed "real" data used to build models. Predictive

419 accuracy will vary depending on the scale of the surveyed data (model inputs) and the spatial

420 extent of the area we seek to predict. Global predictions at high spatial accuracy should

421 therefore be the ultimate goal, which requires increased efforts to standardise and conduct

422 soil biotic and abiotic surveillance at global scales. These advances will be facilitated by better

423 spatial integration of distributed datasets (e.g. global harmonisation of localised climate,

424 geological, remote sensed land cover, and soil datasets) and continued development and

425 validation of mapping predictions against local surveyed data.

431

432 **References**

433 Bivand, R.S., Pebesma,E.J., and Gómez-Rubio,V. 2008. Applied Spatial Data Analysis with

434 R. Springer, ISBN: 978-0-387-78170-9

435 I.R. Booth. 1985. Regulation of cytoplasmic pH in bacteria. Microbiol. Rev. 49: 359–378

436   Bru, D., Ramette, A., Saby, N. P. A., Dequiedt, S., Ranjard, L., Jolivet, C., Arrouays. D.,

437   Philippot, L. (2011). Determinants of the distribution of nitrogen-cycling microbial communities

438   at the landscape scale. ISME J. 5: 532–542.

439   Chapman, D. S., Purse, B. V. 2011. Community versus single-species distribution models for

440   British plants. J. Biogeog. 38: 1524–1535.

441   Cho, J.C., Tiedje, J.M. 2000 Biogeography and degree of endemicity of fluorescent

442   Pseudomonas strains in soil. Appl Environ Microbiol. 66: 5448-56.

443   Comerford, N.B., Franzluebbers, A.J., Stromberger, M.E., Morris, L., Markewitz, D., Moore,

444   R. 2013. Assessment and Evaluation of Soil Ecosystem Services. Soil Horizons, 54

445   Dequiedt, S., Saby, N.P.A., Lelievre, M., Jolivet, C., Thioulouse, J., et al. 2011.

446   Biogeographical patterns of soil molecular microbial biomass as influenced by soil

447   characteristics and management. Glob Ecol Biogeogr 20: 641–652.

448   Dequiedt, S., Thioulouse, J., Jolivet, C., Saby, N. P.A., Lelievre, M., Maron, P.-A., Martin, M.

449   P., Prévost-Bouré, N. C., Toutain, B., Arrouays, D., Lemanceau, P. and Ranjard, L. (2009),

450   Biogeographical patterns of soil bacterial communities. Env. Microbiol. Rep., 1: 251–255.

451   Elith, J., Graham, C.H., Anderson, R.P. et al. 2006. Novel methods improve prediction of

452   species' distributions from occurrence data. Ecography, 29, 129–151.

453   Elton, C.S. 1927. Animal Ecology. University of Chicago Press, Chicago, IL.

454   Fierer N., Jackson, R.B. 2006. The diversity and biogeography of soil bacterial communities.

455   Proc Natl Acad Sci U S A 103: 626–631.

456   Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S.,

457   Gilbert, J.A., Wall, D.H. and Caporaso, J.G. 2012. Cross-biome metagenomic analyses of soil

458   microbial communities and their functional attributes. Proc. Natl. Acad. Sci. U. S. A. 109:

459   21390 - 21395

460 Fierer, N., J. Ladau, J.C. Clemente, J.W. Leff, S.M. Owens, K.S. Pollard, R. Knight, J.A.

461 Gilbert, R.L. McCulley. 2013. Reconstructing the Microbial Diversity and Function of

462 PreAgricultural Tallgrass Prairie Soils in the United States. Science, 342: 621-624.

463 Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M.J., Whiteley. A. S. 2011. The

464 bacterial biogeography of British soils. Env. Microbiol. 13:1642–1654.

465 Hengl, T., de Jesus J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., et al. 2014.

466 SoilGrids1km - Global Soil Information Based on Automated Mapping. PLoS ONE 9(8):

467 e105992. doi:10.1371/journal.pone.0105992

468 Henrys, P., Bee, E., Watkins, J., Smith, N., Griffiths, R.I. 2015. Mapping Natural Capital:

469 Optimising the use of national scale datasets. Ecography 38: 632-638

470 Hirsch, P.R., Mauchline, T.H. & Clark, I.M. 2010. Culture-independent molecular techniques

471 for soil microbial ecology. Soil Biol. Biochem., 42, 878-887.

472 Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M. 2002. Improved Culturability of Soil

473 Bacteria and Isolation in Pure Culture of Novel Members of the Divisions Acidobacteria,

474 Actinobacteria, Proteobacteria, and Verrucomicrobia. Appl Environ Microbiol. 68: 2391-2396.

475 Jenny, H. Factors of soil formation: a system of quantitative pedology. New York : McGraw-

476 Hill, 1941. ISBN 0-486-68128-9

477 Lauber, C.L., Hamady, M., Knight, R., Fierer, N. 2009. Pyrosequencing-Based Assessment

478 of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. Appl

479 Environ Microbiol 75: 5111–5120.

480 Lam, N.S. 1983. Spatial Interpolation Methods: A Review. The American Cartographer

481 10:129-50

482   R Core Team (2014). R: A language and environment for statistical computing. R Foundation

483 for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

484   Ramirez, K.S., Leff, J.W., Barberán, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., et

485   al. Biogeographic patterns in below-ground diversity in New York City's Central Park are

486   similar to those observed globally. Proc. Roy. Soc. B. 2014;281:20141988.

487   Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R.,

488   Fierer, N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil.

489   ISME J. 4:1340-51

490   Spurgeon, D.J., Keith, A.M., Schmidt, O., Lammertsma, D.R., Faber, J.H. 2013. Land-use

491   and land-management change: relationships with earthworm and fungi communities and soil

492   structural properties. BMC Ecol 13:46. doi:10.1186/1472-6785-13-46

493   Stone, D., Blomkvist, P., Bohse Hendriksen, N., Bonkowski, M., Bracht Jørgensen, H.,

494   Carvalho, F., Dunbar, M.B., Gardi, C., Geisen, S., Griffiths, R., Hug, A.S., Jensen, J., Mendes,

495   S., Morais, P.V., Plassart, P., Römbke, J., Rutgers, M., Schmelz, R. M., Sousa, 527 J.P.,

496   Suhadolc, M., Winding, A., Creamer, R.E. 2015. Establishing a Transect for Biodiversity and

497   Ecosystem Function Monitoring Across Europe. App Soil Ecology (this issue)

498   Thomson B. C., Ostle N. J., McNamara N. P., Oakley S., Whiteley A. S., Bailey M. J., Griffiths,

499   R.I. 2013. Plant soil interactions alter carbon cycling in an upland grassland soil. Front.

500   Microbiol. 4:253. doi : 10.3389/fmicb.2013.00253

501   Tóth, G., Jones, A., Montanarella, L. (eds.) 2013. LUCAS Topsoil Survey. Methodology, data

502   and results. JRC Technical Reports. Luxembourg. Publications Office of the European Union,

503   EUR26102 – Scientific and Technical Research series – ISSN 1831-9424 (online); ISBN 978-

504   92-79-32542-7; doi: 10.2788/97922

505   Urich T., Lanzén A., Qi J, Huson D.H., Schleper C., Schuster S.C. 2008. Simultaneous

506   Assessment of Soil Microbial Community Structure and Function through Analysis of the

507   Meta-Transcriptome. PLoS ONE 3(6): e2527. doi:10.1371/journal.pone.0002527

508   Zhou, J., Xia, B., Treves, D. S., Wu, L.-Y., Marsh, T. L., O'Neill, R. V., Palumbo, A.V., Tiedje,

509   J. M. 2002. Spatial and Resource Factors Influencing High Microbial Diversity in Soil. App.

510   Env. Microbiol. 68: 326–334.

511

512

513

514

515

516

517

518

519

520

521

Soil Property	r^2	p	
pH	0.9054	0.001	***
Clay (%)	0.4173	0.001	***
organic C_N_ratio	0.3596	0.001	***
Bulk density (g/cm3)	0.3501	0.001	***
moisture (ml/g)	0.2681	0.001	***
Organic C (%)	0.2624	0.001	***
WHC (ml/g)	0.2521	0.001	***
C (%)	0.2316	0.001	***
Total C_N ratio	0.1961	0.003	**
Silt (%)	0.1372	0.004	**
N (%)	0.1255	0.009	**
Sand (%)	0.0405	0.202	

Table 1. Relationships between soil microbial community ordination axis scores and soil physicochemical properties. Correlations between the NMDS ordination and environmental variables are denoted by r^2 values. Significance (p) was determined by 999 permutations.

Figure 1. NMDS ordination of soil bacterial communities sampled across Europe. Soil pH was found to be the best linear fit to the NMDS ordination scores, and is identified in the plots by a colour gradient denoting the pH for each sample. Centroids are also shown representing the mean score per land use type.

Figure 2. Relationship between soil pH and bacterial community NMDS first axis ordination scores from the 76 sampled soils. The second-degree polynomial fit is also displayed together with 95% prediction intervals.

Figure 3. Using the pH-NMDS1 model determined from 76 soils across Europe to predict NMDS1 scores from >1000 soil pH measurements across Britain. The value for the predictions is shown on the y axis, whereas the x axis denotes the actual observed bacterial community scores from the study of Griffiths et al, 2011. The line shows the fitted least squares regression between the observed and predicted values (R^2 = 0.8).

Figure 4. Interpolated map showing predicted bacterial community ordination scores across EU member states. Colour scale indicates predicted first axis NMDS scores, with negative

scores indicating acidic soils (bogs, acid grassland, upland woods etc) and positive scores indicating communities from more neutral pH soils (productive grassland, arable etc)

Figure 5. Relationship between NMDS first axis scores for bacterial communities and univariate indices of diversity (line denotes a loess fit). Increases in NMDS scores are generally indicative of an increase in bacterial diversity.

Figure 6. Validating spatially mapped predictions of bacterial NMDS scores against national scale survey data for Britain. In all plots the observed data are the actual community scores reported from over 1000 soils sampled across Great Britain (Griffiths et al 2011). Predicted community scores are based on the modelled relationships between bacterial communities and pH from the EcoFINDERs transect sampling fitted to different spatial estimates of soil pH: a) spatial interpolation of community scores predicted from EU wide point data on soil pH (this study); b) geostatistical predictions of topsoil pH values at the global scale (soilgrids.org, Hengl et al, 2014); c) geostatistical predictions of topsoil pH values at the national scale (ukso.org, and see Henrys et al; 2014). Solid red lines show the fitted least squares regression between the observed and predicted values (with associated R^2 displayed in the top right of each plot); and dashed lines display loess fits to illustrate deviations from the linear fit.

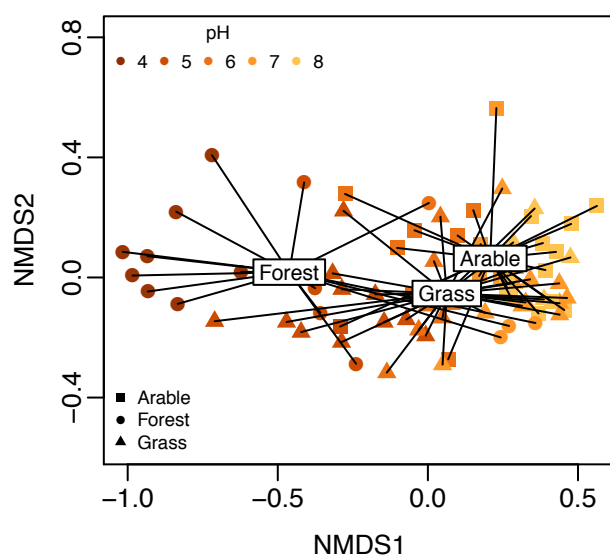**Figure 1.** NMDS ordination of soil bacterial communities sampled across Europe.

Figure 2. Relationship between soil pH and bacterial community NMDS first axis ordination scores from the 77 sampled soils.
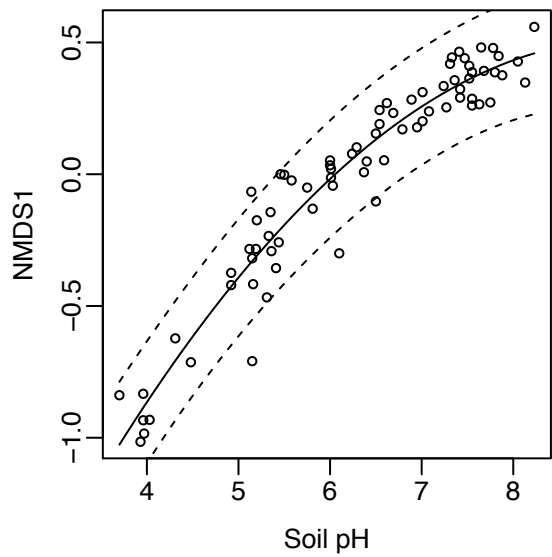
Figure 3. Using the pH-NMDS1 model determined from 77 soils across Europe to predict NMDS1 scores from >1000 soil pH measurements across Britain.
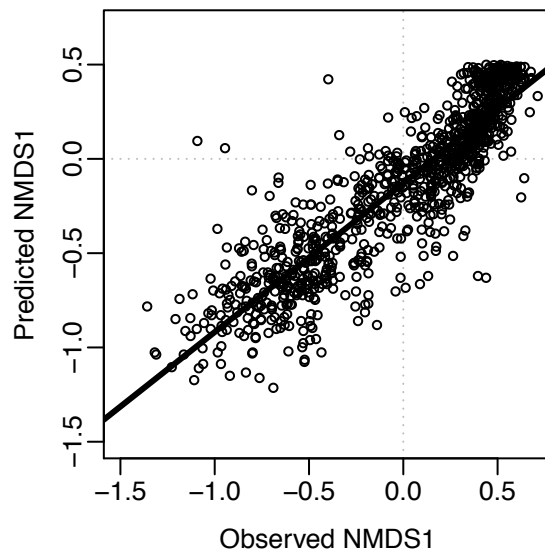
Figure 4. Interpolated map showing predicted bacterial community ordination scores across EU member states.
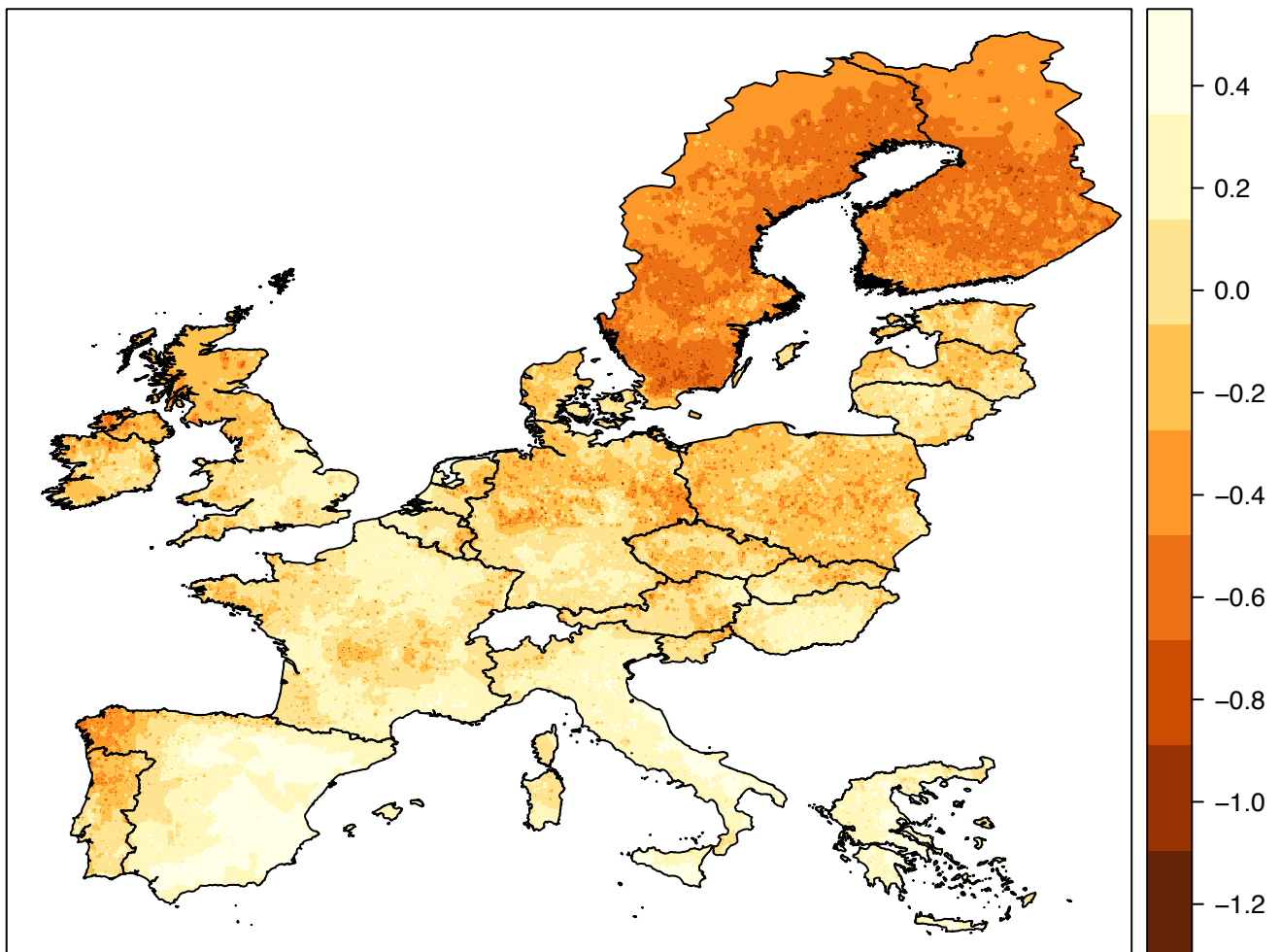
Figure 5. Relationship between NMDS first axis scores for bacterial communities and univariate indices of diversity
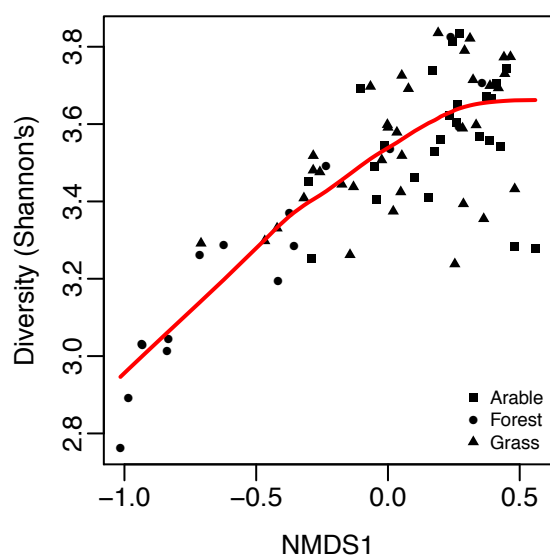
Figure 6. Validating the spatially mapped predictions of NMDS scores against national scale survey data for Britain.