ACCEPTED VERSION

Craig Liddicoat, Peng Bi, Michelle Waycott, John Glover, Andrew J. Lowe, Philip Weinstein
Landscape biodiversity correlates with respiratory health in Australia
Journal of Environmental Management, 2018; 206:113-122 (Includes Supplementary Material, pp. 1-31)

Final publication at http://dx.doi.org/10.1016/j.jenvman.2017.10.007

Landscape biodiversity correlates with respiratory health in Australia

Copy of Final Manuscript, Accepted 6 October 2017
Published online: http://www.sciencedirect.com/science/article/pii/S0301479717309854

Authors: Craig Liddicoat[a,b,*], Peng Bi[c], Michelle Waycott[a,b], John Glover[d], Andrew J. Lowe[a], Philip Weinstein[a]

Affiliations:

[a]School of Biological Sciences and The Environment Institute, The University of Adelaide, North Terrace, Adelaide SA 5005, Australia.

[b]Department of Environment, Water and Natural Resources, GPO Box 1047, Adelaide SA 5001, Australia.

[c]School of Public Health, The University of Adelaide, North Terrace, Adelaide SA 5005, Australia.

[d]Public Health Information Development Unit, Torrens University Australia, Level 1, 200 Victoria Square, Adelaide SA 5000, Australia.

*Author for correspondence: Craig Liddicoat. e-mail: craig.liddicoat@adelaide.edu.au

Keywords: biodiversity, ecological epidemiology, biodiversity hypothesis, lasso

Corresponding author: Craig Liddicoat, School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide SA 5005, Australia, +61 438 843 675, craig.liddicoat@adelaide.edu.au

Abstract

Megatrends of urbanisation and reducing contact with natural environments may pose a largely unappreciated risk to human health, particularly in children, through declining normal (healthy) immunomodulatory environmental exposures. On the other hand, building knowledge of connections between environments, biodiversity and human health may offer new integrated ways of addressing global challenges of rising population health costs and declining biodiversity. In this study we are motivated to build insight and provide context and priority for emerging research into potential protective (e.g. immunomodulatory) environmental exposures. We use respiratory health as a test case to explore whether some types and qualities of environment may be more beneficial than others, and how such exposures may compare to known respiratory health influences, via a cross-sectional ecological epidemiology study for the continent of Australia. Using Lasso penalized regression (to interpret key predictors from many candidate variables) and 10-fold cross-validation modelling (to indicate reproducibility and uncertainty), within different socio-geographic settings, our results show surrogate measures of landscape biodiversity correlate with respiratory health, and rank amongst known predictors. A range of possible drivers for this relationship are discussed. Perhaps most novel and interesting of these is the possibility of protective immunomodulatory influence from microbial diversity (suggested by the understudied 'biodiversity hypothesis') and other bioactive agents associated with biodiverse environments. If beneficial influences can be demonstrated from biodiverse environments on immunomodulation and human health, there may be potential to design new cost-effective nature-based health intervention programs to reduce the risk of immune-related disease at a population level. Our approach and findings are also likely to have use in the evaluation of environment and health associations elsewhere.

1. Introduction

There is growing awareness of the numerous mechanisms and co-benefits that link natural and biodiverse environments with human health. It is important to understand such connections given the global challenges of escalating population health costs and declining biodiversity (WHO and SCBD, 2015). As reviewed elsewhere (Craig et al., 2016; Keniger et al., 2013; Myers et al., 2013; Sandifer et al., 2015; WHO and SCBD, 2015), there are broad and interacting mechanisms of human health impact from environmental change, many of which are well-studied. However, some potentially important environmental influences on human health remain understudied due to their multidisciplinary nature. A key example is described by the biodiversity hypothesis (von Hertzen et al., 2011), and related microbial 'old friends' mechanism (Rook, 2013), as recognized by the World Health Organisation (WHO and SCBD, 2015) and World Allergy Organization (Haahtela et al., 2013), which highlights that environmental microbiota (or communities of microorganisms from the surrounding environment) overlap and interact with human commensal microbiota, contribute to human microbial diversity and may provide important beneficial immunomodulatory roles. As discussed later, there are potential links between different environments, their microbiotas and other possible bioactive agents (e.g. volatile organic compounds, VOCs; air ions), and via direct and aerobiological exposures, possible immunomodulatory effects and human health influences (e.g. see Liddicoat et al. (2016) Fig. 1).

The possibility of populations receiving some level of inadvertent ambient beneficial or adverse immunomodulatory influence associated with different types and qualities of environment, highlights a potentially important gap in our awareness of possible links between environments and human health. Megatrends of urbanisation and reducing contact with natural environments may pose a largely unappreciated risk to human health through declining normal (healthy) immunomodulatory environmental exposures. Children may be

73 particularly impacted by inadequate exposures during the critical early period of immune
74 system development (Wopereis et al., 2014) where immunoregulatory commensal microbiota
75 can be acquired through random environmental encounters (Artis, 2008). For example,
76 children who grow up in environments with diverse microbial exposures such as traditional
77 farms, are less prone to developing asthma and atopy (allergic sensitization) (Ege et al.,
78 2011; Stein et al., 2016). Beneficial immunomodulatory environmental exposures are also
79 suggested into adulthood (Douwes et al., 2007; Rottem et al., 2015; von Hertzen and
80 Haahtela, 2006). Moreover, a lack of appropriate environmental exposures and deficient
81 immune training and regulation may also impact other areas of immune-related human health
82 including susceptibility to infectious disease (as below).
83 Respiratory health provides a conspicuous test case to explore environment-human
84 health associations that have plausible microbiota-mediated linkages, because breathing
85 offers a primary mode of exposure for people interacting passively with the environment.
86 Throughout life, millions of litres of air move through the human respiratory tract, which
87 provides one of the first points of contact with environmental contaminants and bioaerosols
88 (airborne microbiota and bioactive agents). Airway epithelial tissues provide a protective
89 arsenal of physical barriers, niche-occupying commensal microbiota, antimicrobial
90 compounds, and receptors ready to orchestrate immune responses (Parker and Prince, 2011;
91 Whitsett and Alenghat, 2014). Environmental microbiota can interact directly with
92 respiratory mucosal immune receptors or via ecological interactions with host commensal
93 microbiota. Similar interactions can also occur in the gut, influencing immune- and health-
94 status, after environmental microbiota deposit in the airways and are transported by cilia to be
95 swallowed (Rook, 2013). Importantly, dysregulation of the airway epithelial innate immune
96 system can be associated with compromised immunity and chronic inflammation (Parker and
97 Prince, 2011). Immune dysfunction may involve adverse feedbacks that reinforce imbalance

98 (or dysbiosis) of host microbiota (Haahtela et al., 2013), which in turn may favour pathogenic
99 microbes and increase susceptibility to infectious disease. As discussed later, there is often
100 not a clear distinction between infectious and non-infectious respiratory disease outcomes, for
101 example, where one type of disease (e.g. cold or influenza) can exacerbate symptoms of
102 another (e.g. asthma). This means there is potential for environments and their microbiota to
103 impact immune status, which in turn may have potential broad underlying (and population-
104 level) influence on multiple infectious and non-infectious respiratory diseases.

105 If particular macro- and landscape-scale features of the environment (e.g. types of
106 vegetation, soil, land use, and their diversity) can be associated with human health benefits
107 (and ultimately supported by new knowledge of underlying causal mechanisms), it may be
108 possible in the future to design new cost-effective, landscape and urban green space
109 interventions with concurrent benefits for public health and biodiversity conservation.
110 Informing such outcomes would require a large body of multidisciplinary research. The work
111 presented here represents an early step.

112 Our motivation for this study is to build insight and provide context and priority for
113 further research into these types of potential beneficial environmental exposures, through
114 building on existing, inexpensive data. Given the possible abovementioned links between
115 environments, immune development or dysfunction, and infectious and non-infectious
116 respiratory disease, we examine available aggregated respiratory health outcome data in a
117 cross-sectional ecological epidemiology study spanning the continent of Australia. Our aim is
118 to test whether some types and qualities of environment may be more beneficial than others,
119 and how such exposures may compare to known respiratory health influences. We appreciate
120 that using aggregated health response data represents a limitation in terms of loss of
121 specificity to link environmental influence with any particular disease. On the other hand, this

approach may offer greater sensitivity to detect possible broad environmental influence on multiple respiratory disease outcomes.

Due to many unknowns (e.g. possible agents, behaviourally- and temporally-mediated exposures, requisite exposures, immunomodulatory and other possible physiological pathways), the use of environmental proxies is warranted as a pragmatic investigation tool (Liddicoat et al., 2016). Proxies allow us to consider a variety of (including possible beneficial immunomodulatory) environmental influences on human health. We might expect to see correlative signals between health outcomes and environmental exposures that are consistent with sources of microbial diversity, such as biodiverse environments, diversity in land use, and soils high in clay and/or organic matter content (Liddicoat et al., 2016; Rook, 2013; von Hertzen and Haahtela, 2006).

We use a data-intensive approach suited to this emerging area of scientific inquiry where it is important to gain an early understanding of key relationships among many variables, and note that models will improve iteratively over time (Elliott et al., 2016). Our modelling approach reflects the highly faceted nature of environments and is adept at handling large numbers of (included potentially correlated) candidate predictors. To guide future work, we provide clear interpretation and ranking of previously unaccounted environmental influences among important predictors of our respiratory health data.

2. Methods

We use an array of specially-prepared environmental covariates to estimate environmental exposures, each allowing for a potential surrounding zone of influence (section 2.2). To cater for geographically-variable environmental and multimodal social predictors, we stratify our analysis into three different socio-geographic groups spanning the Australian continent (section 2.3). We develop an automated screening algorithm to filter out extraneous variables

147 and assess selected transformations of candidate predictors to help optimize linear
148 relationships in subsequent modelling (section 2.4). We use least absolute shrinkage and
149 selection operator (or Lasso) penalized regression (a contemporary machine-learning
150 algorithm (Tibshirani, 1996)) to interpret key predictors from large numbers of candidate
151 variables, and purpose-built 10-fold cross-validation (CV) modelling to indicate
152 reproducibility and uncertainty in our results (section 2.5). This approach puts the onus on
153 variables to compete and display strength and consistency of predictive value.
154

155 **2.1. Public health and contextual data**

156 We use merged 2011/12 and 2012/13 Social Health Atlas of Australia (PHIDU, 2015, 2016)
157 data for respiratory disease public hospital admissions (a breakdown of disease codes is
158 provided in Appendix A, Table S1), together with accompanying contextual data (Appendix
159 A, Table S2). For reliability and privacy purposes, these respiratory disease and associated
160 contextual data are only available in aggregated form. In this study data are aggregated
161 spatially, by Australian local government areas (LGAs), and thematically, grouped under
162 principal diagnoses of diseases of the respiratory system. In each LGA, the 2011-13
163 normalized mean cumulative incidence of respiratory disease public hospital admissions was
164 calculated using:

165 $$ASR_{2011-13} = \left(N_{11/12} + N_{12/13}\right) / \left(\frac{N_{11/12}}{ASR_{11/12}} + \frac{N_{12/13}}{ASR_{12/13}}\right) \qquad (2.1)$$

166 where *ASR* is the respective age standardized rate per 100,000, and *N* is the respective raw
167 number of annual admissions recorded. We used the LGA-based data because we considered
168 they offered a balanced coverage of health reporting areas and a spatial framework suited to
169 capturing environmental variability across both urban and regional Australia. Numbers of
170 LGAs used in the modelling are shown in Table 1.
171

2.2. Environmental data

We prepared and collated an as large as practical number of environmental covariate layers to reflect possible direct and indirect influences on respiratory health. Many environmental layers were included to allow detection of as-yet-unexplained possible microbiota-mediated and other influences, as introduced earlier and discussed elsewhere (Liddicoat et al., 2016; Rook, 2013). Our data-intensive approach (i.e. including many variables and later use of Lasso machine-learning) also aimed to reduce modelling bias that might arise through pre-selection of only a small number of candidate environmental variables. We represent the diversity, and multi-faceted nature of environments using an array of climatic, soil, landscape and vegetation-based variables based on exhaustive Australia-wide gridded mapping datasets (Appendix A, Table S3). For consistent and pragmatic data-handling we adopt a common 250 m resolution grid system, as used elsewhere in continent-wide, landscape-scale mapping of land cover (Geoscience Australia, 2014). Where necessary to match the common grid system, resampling was performed bilinearly for numeric data layers, and using the nearest neighbour method for categorical layers.

Environmental map layers were then re-expressed using focal neighbourhood statistic calculations so that environmental data (at any cell location, or if averaged over an area) provide an estimate of exposure corresponding to a surrounding zone of influence. Such a zone reflects potential movement of populations within their surroundings and also the possibility of airborne dispersal of environmental microbiota and bioactive agents. We do not know how far populations or bioaerosols disperse, however we chose a nominal 3 km radius area as our representative environmental zone of influence. This area was consistent with previous studies examining possible links between land use and human immunomodulatory influence (Hanski et al., 2012) and green space and self-reported health (Maas et al., 2006). We were also guided by Ruokolainen et al. (2015), who found the spatial scale of land-use

description affected the detection of statistically significant relationships between land-use and atopy (allergic sensitization), which they observed in the range 2-5 km.

This meant for all of the Australia-wide gridded environmental data layers, in every grid cell, we calculated a measure summarising a particular aspect of the surrounding environment over a 3 km-radius area. Numeric variables were averaged (including climatic, soil, landscape, and remotely-sensed vegetation parameters), while class proportions and Shannon diversity indices were calculated for categorical themes (i.e. land use, land cover, ecological land units and major vegetation groups). The conversion of categorical data to numeric data (using the focal neighbourhood calculations) also enabled a simpler linear modelling approach as all environmental data were ultimately expressed in numeric form. Additional information and formulae for the calculation of environmental layers are provided in Appendix A. A total of 176 gridded environmental-variable map layers were prepared.

To join with available health outcome data in LGAs, environmental data were averaged within each LGA boundary. (For future studies where finer resolution health data are available, we suggest these specially-prepared focal neighbourhood environmental layers could be sampled at point locations or averaged over smaller areas.)

Point-based air pollution data were also considered, however these were handled differently to the gridded data layers (due to reasons discussed below). Estimated total industrial emissions of inhalable particulate matter (of 10 micrometres or less in diameter, or PM10) for 2011/12 and 2012/13 were spatially intersected and summed in each LGA, expressed as area-based rates ($kg.km^{-2}.yr^{-1}$), then averaged to estimate the mean 2011-13 emissions.

220 **2.3. Socio-geographic clustering**

221 Environments vary greatly across the Australian continent (approx. area of 7.7 million km^2),
222 with populations spanning urban, peri-urban, rural and remote locations. Therefore, we might
223 expect any environmental influences on health (where present) to vary geographically. There
224 are known health inequalities associated with socioeconomic status, remoteness (AIHW,
225 2007), and Aboriginal populations (Gubhaju et al., 2013); and these social factors also vary
226 geographically. During preliminary data analysis we observed a negative skew in the
227 Australia-wide distribution of Socioeconomic index, suggestive of a second minor mode of
228 lower socioeconomic status LGAs (Appendix A, Fig. S1). To allow for likely varying
229 environmental and multimodal socio-geographic influences across an expansive dataset, and
230 to improve linear modelling of the respiratory health outcome, we stratify the data into three
231 socio-geographic groups using k-means clustering based on all available data for
232 Socioeconomic index, Population density and Percent Aboriginal persons (Fig. 1; Appendix
233 A, Table S4). We broadly interpret the resulting groups (LGA clusters) as the 'Moderate
234 majority', 'Major cities', and 'Remote disadvantaged'. Separate analyses were performed
235 using these socio-geographic clusters.
236

237 **2.4. Additional data preparation for modelling**

238 Due to missing data, we excluded some LGAs from the health response modelling. We also
239 developed an automated screening algorithm to objectively exclude extraneous variables (e.g.
240 irrelevant environment class proportions in particular LGA clusters), and to consider a
241 limited set of candidate predictor transformations (designed to improve linear relationships in
242 subsequent modelling). The screening algorithm was fully coded in R script as a means to
243 transparently and consistently inspect variables and consider preparatory steps that might
244 otherwise be done manually one variable at a time, prior to multiple linear regression

modelling. In each LGA cluster, pragmatic threshold criteria were used to exclude (original or transformed) variables if less than 50% of LGAs contained non-zero values, if variable skewness exceeded 1.5, kurtosis exceeded 4, or if stand-alone explanatory value (R^2) was less than 2.5%. Only logit transformations were considered for proportion or percentage data, otherwise square root, log10, square and cube root were considered, subject to certain disqualifications (e.g. log10 cannot be used on zero or negative values). Selection of a final representative candidate (whether original or transformed) was made according to pre-determined rules that attempted to balance the trade-off between interpretability and optimising normality of predictor distributions. Further description of the data preparation steps are provided in Appendix A (including R code).

Within LGA clusters, all screening and subsequent analysis of candidate predictors was performed on 95% Winsorized data, designed to objectively eliminate the influence of extreme and potentially outlying values (Friedman and Popescu, 2008). The distributions of respective health response variables were inspected, and for the 'Moderate majority' and 'Remote disadvantaged' clusters, log10 variance-stabilising transformations were applied (Appendix A, Fig. S2). All predictor data were centred and scaled before input for Lasso modelling.

2.5. Lasso modelling

We use Lasso penalized regression as our primary tool for correlation analysis and health response modelling, as implemented in the R *glmnet* package (Friedman et al., 2010). Use of the Lasso is relatively new in epidemiology, but has been suggested as a credible alternative to more conventional stepwise multiple regression approaches when identifying key predictors from large datasets (Mansiaux and Carrat, 2014). As the Lasso can be prone to inconsistency in variable selection (Leng et al., 2006), within each LGA cluster, we apply

explicit randomized 10-fold resampling to generate variation in input data for modelling. This was designed to help interpret the levels of variation, reproducibility and uncertainty in our modelling results. In each fold, internally within the Lasso software we ran leave-one-out CV to identify the optimal penalisation parameter (we used the parsimonious option corresponding to s = 'lambda.1se' when calling the *cv.glmnet* function, with the Lasso model corresponding to the default setting a=1). Details of the statistical analyses and R scripts used for the modelling are provided in Appendix A.

We interpret the relative importance and direction of predictors identified by the Lasso from the size and sign of standardized regression coefficients (this is facilitated by centring and scaling of data as discussed). Identified predictors and their standardized coefficients were harvested from the respective 10-fold Lasso modelling outputs and plotted, as below.

3. Results and Discussion

3.1. Key predictors of respiratory health

We show results for the 'Moderate majority' in Fig. 2, ordered by the mean absolute size of standardized regression coefficients. Important predictors (toward the top of Fig. 2) are consistently identified. Beneficial respiratory health outcomes were associated with (in order of decreasing importance) Socioeconomic index, Diversity of major vegetation groups, Species richness (log10), Proportion of eucalypt forests 10-30m (logit), Percent overweight persons (logit), Percent English-speaking immigrants (logit), Proportion of open [i.e. 30-70% canopy cover] trees (logit), Diversity of land use, and Proportion of nature conservation (logit). Adverse respiratory health associations were identified with Distance to coast, Percent obese persons, Mean temperature annual range, Maximum temperature of warmest month, Percent smoking during pregnancy, Proportion of warm wet plains (logit), Percent Aboriginal

295 persons (logit), and Vegetation fractional cover minimum nonphotosynthetic (logit).
296 Predictors decrease in importance down the y-axis and we suggest variables towards the
297 bottom should be viewed with caution (e.g. Mean precipitation of the coldest quarter,
298 Vegetation fractional cover minimum photosynthetic (logit), and Soil cation exchange
299 capacity * erodible fraction (geometric mean)). Predictors that are rarely or not consistently
300 selected may not be generally applicable (e.g. localized influence) and, possibly, the lowest
301 ranked variables may be spurious (e.g. analogous to near-zero coefficient variables in a
302 multiple linear regression). Many variables were not selected at all by the Lasso in any of the
303 10 folds (see Appendix A, Table S5).

304 The nature of correlations between predictors identified across the 10-fold Lasso
305 modelling is shown, by way of example, for the 'Moderate majority' cluster (Appendix A,
306 Fig. S3). Important predictors for the remaining clusters are shown in Appendix A, Fig. S4-
307 S5. Summary performance statistics (Table 1) and CV plots (Appendix A, Fig. S6) for the 10-
308 fold Lasso modelling, indicate that moderate levels of prediction success were achieved.
309 Based on concordance correlation coefficients (which indicate how closely observed and
310 predicted values adhere to a 1:1 relationship), moderate agreement was found between
311 predicted and observed health responses in the 'Moderate majority' and 'Major cities'.
312 However, poorer agreement was found for the 'Remote disadvantaged' LGAs, possibly due
313 to their sparsely settled nature, coarse scale of data and small sample size. The high mean R^2
314 value for the 'Remote disadvantaged' cluster (Table 1) suggests good linear agreement within
315 each of the CV folds. However, we view this with caution due to small sample numbers and
316 poor 1:1 alignment of observed and predicted values indicated by the low concordance
317 correlation coefficient and Fig. S6 (c) (Appendix A).

318 Our approach provides a side-by-side comparison of the potential influence of
319 (including previously unaccounted) environmental variables against recognized population

health predictors of respiratory disease (Fig. 2). Reassuringly, a number of key predictors in our results match expectation, or align with findings elsewhere for lifestyle and environmental influences. For example, higher socioeconomic status (Socioeconomic index) associates with reduced hospital admissions and smoking associates with increased hospital admissions. Noting that our dependent variable is aggregated, we draw tentative support from literature relating to both general and disease-specific influences linked to respiratory health outcomes. For example, obesity has been negatively associated with respiratory health (Zammit et al., 2010), and lesser indications of a counter-intuitive health advantage attributed to being overweight are not without precedent (Flegal and Kalantar-Zadeh, 2013). Benefits from closeness to sea air (or disadvantage with distance from coast) might be expected with a number of respiratory conditions, for example in non-cystic fibrosis bronchiectasis (Kellett and Robert, 2011). Although, the variable of Distance to coast provides an example for possible parallel interpretations, such as respiratory conditions that could be linked to increasing dust in drier inland areas. High temperatures are reported to have an impact on respiratory admissions, particularly in the elderly (Michelozzi et al., 2009). Warm and wet (humid) environments also contribute to population risk of non-tuberculous mycobacterial pulmonary infection (Prevots and Marras, 2015).

Our results highlight a number of health-correlated environmental variables that are worthy of further investigation (refer to beneficial respiratory health associations listed earlier). In particular, Diversity of major vegetation groups featured in our results—and provides a measure of differentiation among habitats (analogous to beta diversity). The higher ranking of biodiversity surrogates (e.g. Diversity of major vegetation groups, Species richness (log10), Proportion of eucalypt forests 10-30 m (logit), Proportion of nature conservation), compared to remote sensing of vegetation greenness (e.g. fraction of photosynthetically active radiation and fractional cover layers) and land use class proportions

345 for broad-acre agricultural uses (which were considered but not recognized in the modelling),
346 is consistent with the notion that the quality of environments is important to respiratory
347 health. From this, we speculate that optimal health benefit may not be merely associated with
348 any type of green space and 'clean country air', but may be promoted by as-yet-unknown
349 attributes of the green space itself. However, air pollution has not been fully accounted for, as
350 discussed below. We also undertook extended analyses for the 'Moderate majority' LGA
351 cluster to better understand the significance of variables selected by the Lasso (noting this
352 represents an evolving area of statistical science), and this is described in the Appendix A.

353 We saw a notably different pattern for 'Major cities' (Appendix A, Fig. S4). Health
354 benefits were primarily associated with socioeconomic status (consistent with the notion of
355 social 'insulating layers' (Myers et al., 2013)), while other social, lifestyle, and ambient
356 environmental influences appeared to be generally detractive. For example, possibly, rainfed
357 pasture and peak levels of remotely-sensed living vegetation could in this case be related to
358 excessive levels of airway allergens from productive but low-biodiversity neighbouring
359 farmland. It is likely the scale of our data has limited the prospects of detecting positive
360 environment-health associations in this cluster.

361 The 'Remote disadvantaged' cluster was only represented by a small and coarse-scale
362 dataset, so results may be misleading, however it was interesting to see wetlands, *Acacia*
363 forests and woodlands, and swampy vegetation feature positively—also consistent with the
364 notion that biodiversity may provide an as-yet-undetermined beneficial influence on
365 respiratory health.

366 Previous modelling of Australia-wide asthma and chronic obstructive pulmonary
367 disease (COPD) hospitalisation rates (AIHW et al., 2014) found significant associations with
368 socioeconomic status, remoteness and the Indigenous proportion of the population; with

generally higher hospitalisation rates for both asthma and COPD in inland and rural Australia. In general terms, our findings are consistent with previous studies examining environmental influences on respiratory health. However our approach in considering an array of environmental attributes (including landscape-scale biodiversity surrogates), provides potential new insight and priority research targets to further investigate causal mechanisms.

3.2. Air pollution

We lacked appropriate exhaustive, continent-wide mapping data to fully represent the potential influence of air pollution in this study. We did test the influence of point-based mean 2011-13 estimated total industrial PM10 emissions in each LGA, however these data did not register as a key predictor in our analyses. Closer inspection of the PM10 data showed many LGAs contained zero values which made the data poorly distributed in the context of our modelling approach (due to inherent skewness and also disqualification of log10 transformation due to zeros). We note these PM10 data represent modelled estimates (not actual measured data) and don't account for natural emissions (e.g. from windblown dust, sea salt aerosols and biological aerosol particles), which elsewhere (Liora et al., 2015) are estimated to be of a similar order to the lower range of estimated industrial PM10 emissions analysed here. Without wider data or knowledge of natural background PM10 levels we did not consider it valid for this study to interpolate a continent-wide map layer based on incomplete and spatially-limited data.

We suggest the lack of appropriate air pollution data doesn't invalidate our broader results; for example, we are encouraged by the recognition of known respiratory health predictors. Also, Australia's ambient air quality is generally good, typically meeting national

health-based standards, with many key air pollutants having declined or remaining stable over the assessment period 1999-2008 (State of the Environment 2011 Committee, 2011). Particulate pollution from localized extreme events such as bushfires has been linked to increased respiratory hospital admissions (Chen et al., 2006), although bushfire frequency mapping was included but not identified as a useful predictor in our study. Urban air pollution is another localized concern for respiratory health (Simpson et al., 2005) that is not accounted for due to a lack of exhaustive data. To some extent, the separation of 'Major cities' may largely quarantine this factor in our analyses. We expect air pollution may contribute to respiratory health outcomes, however our data did not support findings of an association in this study.

3.3. Possible links between biodiverse environments and respiratory health

Our finding of an association between beneficial respiratory health outcomes and surrogate measures of landscape biodiversity (in particular, Diversity of major vegetation groups, and other measures listed earlier) warrants consideration of possible underlying ecological linkage mechanisms. As a megadiverse continent (Groombridge, 1992), Australia provides an interesting study area for exploring such links.

Simple indirect explanations may exist, for example, due to absence of air pollution (or 'clean country air'). Seasonal exposures to airborne airway allergens such as grass pollen, a major trigger for allergic rhinitis and asthma (Davies et al., 2015), may be indirectly related to landscape biodiversity. For example, where landscapes and land uses elsewhere may be characterized by a low biodiversity of plant species and a corresponding concentration of (often introduced) grass and tree species producing large amounts of allergenic pollen (Cariñanos and Casares-Porcel, 2011). In Australia, knowledge of seasonal and

417 biogeographical patterns of pollen exposures and relationships with allergic respiratory
418 disease is growing (Davies et al., 2015). Human health results from a complex interplay of
419 factors, so a number of (including environmental) influences may possibly contribute. Viral
420 infections are known to exacerbate other respiratory conditions, and are implicated in 50 to
421 80% of all hospitalisations for asthma (Bardin, 2004). This highlights the importance of
422 underlying immune status and possible connections between common communicable and
423 non-communicable respiratory disease.

424 In terms of general health outcomes, cross-sectional spatial epidemiology studies have
425 previously found positive associations between neighbourhood green space and self-reported
426 health status (e.g. de Vries et al., 2003; Maas et al., 2006). Exposure to green space and
427 natural environments has been associated with reduced all-cause mortality (Mitchell and
428 Popham, 2008). Mitchell and Popham also found that increased exposure to green space
429 reduced health inequalities related to socioeconomic status. Common beneficial influences
430 suggested from enhanced visual and/or physical access to green space include greater
431 physical activity, reduced stress, enhanced restoration from mental fatigue, improved mood
432 and self-esteem, and provision of psychological and mental health benefits (reviewed
433 elsewhere e.g. (Keniger et al., 2013; WHO and SCBD, 2015)). These influences may act
434 synergistically to further enhance benefits from 'green exercise' (exercise in natural
435 surroundings) (Gladwell et al., 2013).

436 A number of reviews also suggest the possibility of beneficial environmental
437 microbiota- and bioactive-mediated immunomodulatory influences on human health (Craig et
438 al., 2016; Haahtela et al., 2013; Rook, 2013; Sandifer et al., 2015; WHO and SCBD, 2015).
439 Intuitively, there may be plausible mechanisms for microbiome-respiratory system health
440 influence including competitive exclusion and/or regulation of airway pathogens and immune
441 priming. There is a known role for host commensal microbiota in normal (healthy) immune

functioning (e.g. in the gut (Artis, 2008; Molloy et al., 2012)) and increasing awareness of associations between dysbiosis (i.e. altered composition of the human microbiota, often with an imbalance between pathogenic and commensal organisms) and a range of diseases (Belizario and Napolitano, 2015; Clemente et al., 2012), including respiratory diseases (Dickson et al., 2013; Noval Rivas et al., 2016). Certain commensals, particularly in the gut, may play important roles in the signalling and control of inflammatory and regulatory immune processes (Artis, 2008; Molloy et al., 2012). Emerging knowledge of a 'gut-lung' microbial axis (Fujimura and Lynch, 2015; Noval Rivas et al., 2016), or connection between host gut and airway microbiota, has been demonstrated in mouse models where changes in gut microbiota composition can influence airway immune responses (Fujimura et al., 2014; Ichinohe et al., 2011; Kim et al., 2014).

Importantly, at least a portion of the human microbiota is in dynamic exchange with environmental microbiota and, therefore, natural microbial diversity is recognized as an important potential contributor to healthy immune functioning (Haahtela et al., 2013; WHO and SCBD, 2015). Aerobiology studies tell us that the air is alive with all manner of biogenic and bioactive particles (Després et al., 2012; Polymenakou, 2012). Bioaerosols often comprise microorganisms from soils, plant surfaces, water bodies, rocks and built structures that are mobilized by wind and splashing water and can be transported considerable distances (Polymenakou, 2012). Bioaerosols can be deposited in the upper airways, carried up the trachea by the action of cilia and then swallowed—therefore, airborne microorganisms can end up on the skin, in the airways, and in the gut where they can perform immunomodulatory roles (Rook, 2013). Pulmonary neuroendocrine cells in the airway can also directly translate environmental cues into physiological and immune responses (Branchfield et al., 2016). Possible connections between environmental features (e.g. vegetation, soils, water bodies, land use, etc.) and environmental microbiota are discussed elsewhere (Liddicoat et al., 2016;

467 Rook, 2013), and such knowledge is steadily building (e.g. via Bissett et al., 2016; Gilbert et
468 al., 2014). Neighbouring land use and landscape composition have been shown to influence
469 bioaerosol composition (Bowers et al., 2011; Després et al., 2012; Mhuireach et al., 2016).
470 However, more work is needed to investigate possible links between aerobiological
471 exposures and surrounding environments.

472 Although not explored here, temporal aspects of environmental microbiota-human
473 contact are likely to be important, in relation to commensal microbiota and immune system
474 development (Wopereis et al., 2014). In immature hosts, commensals appear to be acquired
475 through random environmental exposures (Artis, 2008). Once established, the mature host
476 commensal microbiota is more resilient to colonisation (Seedorf et al., 2014). Although, if we
477 extrapolate from studies of short- versus long-term dietary change (Voreades et al., 2014),
478 possibly, long-lasting environmental change may influence a shift in composition even for
479 mature commensal microbiota. Poor western diets may cause irreversible loss of potential
480 key beneficial commensal taxa, with suggestions that a rewilding of the human microbiota
481 may be needed (Sonnenburg et al., 2016). Recent findings that the human intestinal
482 microbiota may be dominated by spore-forming bacteria (Browne et al., 2016) which are
483 designed to persist in the environment, aligns with the idea that critical early and perhaps
484 'ongoing maintenance' environmental exposures may help replenish key spore-forming
485 bacteria to the human commensal microbiota. In turn, as discussed, such interactions may
486 influence airway immune status and respiratory health.

487 Health benefits may also be linked to abiotic bioactive agents from the environment
488 such as phytoncides (wood essential oils or VOCs) and negative air ions (Craig et al., 2016).
489 Negative air ions from waterfall aerosol have been shown to produce lasting beneficial
490 effects on asthma symptoms, lung function and airway inflammation (Gaisberger et al.,
491 2012). Plants and phyllosphere microbiota are known to emit and influence a wide variety of

492 VOCs (Bulgarelli et al., 2013). Such plant-based VOCs can promote or inhibit (and thus
493 shape) adjacent microbial communities (Bringel and Couée, 2015), which may in turn have
494 varying human health influences.

495 Australia has among the highest rates of allergies among developed countries,
496 affecting almost 20% of the population, and this prevalence is increasing (Access Economics,
497 2007). According to the 2014-15 National Health Survey (ABS, 2015), long-term respiratory
498 disease is estimated to affect nearly 31% of the population; with the prevalence of hayfever
499 and allergic rhinitis, asthma, and chronic sinusitis estimated at 19.4%, 10.8%, and 8.4%
500 respectively. If links between macro- and landscape-scale environmental biodiversity and
501 beneficial immunomodulatory influences can be substantiated and quantified, significant
502 reductions in disease burden and public health expenditure might be possible through greater
503 consideration of biodiversity in public health programs and new landscape and urban green
504 space design.

505

506 **3.4. Limitations**

507 The environmental mapping products used here have been produced from a variety of
508 projects and methods, independently from routine Australia-wide public health reporting.
509 Consequently, the trade-off for representing a wide diversity of potential environmental
510 influences, is a loss of exact temporal coincidence of surrogate environmental exposures
511 (derived from environmental mapping) with the health response data. We continued on the
512 basis that the rate of environmental change across the continent should be sufficiently slow,
513 or health impacts possibly integrated over time, that available environmental mapping
514 datasets from recent decades should sufficiently characterize the type of environments and
515 potential exposures relevant to our recent (2011-2013) health response data. We note this
516 approach allows us to consider many expert-interpreted mapping layers for different

517 classifications of the environment. Our approach is vindicated where a number of these
518 expert-interpreted layers (and derived biodiversity surrogates) offer greater explanatory value
519 than less-interpreted remote sensing-derived layers (e.g. based on greenness indices) also
520 included in the modelling.

521 We recognize there will be environmental and social/lifestyle influences, and
522 potential lurking variables and interactions, that we have not specifically accounted for (e.g.
523 inadequate air pollution data, as discussed). Ecological epidemiology studies will always
524 contain confounders, and cross-sectional studies are generally limited as they cannot infer
525 temporal directionality or causality. However this type of study is suited to inexpensive
526 examination of population-wide disease associations with multiple potential influences, as
527 well as further hypothesis generation and informing more detailed studies. As our public
528 health data in this study were aggregated into LGAs we note the potential for erroneous
529 findings due to ecological bias and fallacy (Elliot et al., 2000), and the modifiable areal unit
530 problem (Parenteau and Sawada, 2011). The scale of data is generally finest to coarsest in
531 order from 'Major cities', 'Moderate majority', then 'Remote disadvantaged'—where coarser
532 datasets are more prone to errors of the type suggested. To some extent, the use of 10-fold
533 resampling and CV modelling, moderates concern due to the modifiable areal unit problem,
534 particularly in 'Moderate majority' cluster, where top-ranking predictors are consistently
535 identified from different (10-fold) sub-selections of LGAs.

536 The use of aggregated respiratory disease outcome data poses a limitation for
537 interpretation of possible specific environment-respiratory disease relationships. On the other
538 hand, our approach may offer insights to possible broad underlying influences (through as-
539 yet-unknown mechanisms) that may impact a range of associated diseases. Also we note that
540 the use of aggregated health outcome data is common in population health studies (e.g. (de
541 Vries et al., 2003; Maas et al., 2006; Mitchell and Popham, 2008)). On the positive side, with

this inexpensive but expansive continent-wide dataset, we have provided a side-by-side comparison of known population-level respiratory health predictors and previously untested potential environmental influences.

4. Conclusions

Using continent-wide datasets, we have explored previously unaccounted potential environmental influences on respiratory health. Our results are generally suggestive of a potential beneficial or protective health influence associated with natural and biodiverse environments. Across different socio-geographic settings, we highlight environmental features and attributes worthy of more targeted investigation—in particular, high biodiversity areas, natural forests through to open woodlands, wetlands, high diversity in land use, and proximity to coastal areas. The validity of these findings is supported by the parallel associations we found with well-recognized social predictors of health included in the modelling. These findings provide additional motivation to investigate the various potential connections between biodiverse environments and human health. Among these, we suggest possible beneficial immunomodulatory influences from environmental microbiota and bioactive agents (perhaps associated with various environmental components such as types of vegetation, soils, land use, and their diversity) represent a worthy research focus (e.g. via study of environmental and human microbiomes and human health biomarkers). Our results provide additional support and context to emerging research into the potential role of natural green space exposures in reducing the risk of immune-related disease at a population level. Our approach can be readily adapted to explore potential health and environmental associations elsewhere.

Acknowledgements. We thank Graham Rook and Geraint Rogers for helpful discussions and suggestions. This research used high-performance computing services provided by eRSA (www.ersa.edu.au), and mapping datasets produced through Australia's Terrestrial Ecosystem Research Network (TERN, www.tern.org.au).

Appendix A. Supplementary data

Supplementary material related to this article can be found at <insert Science Direct link>

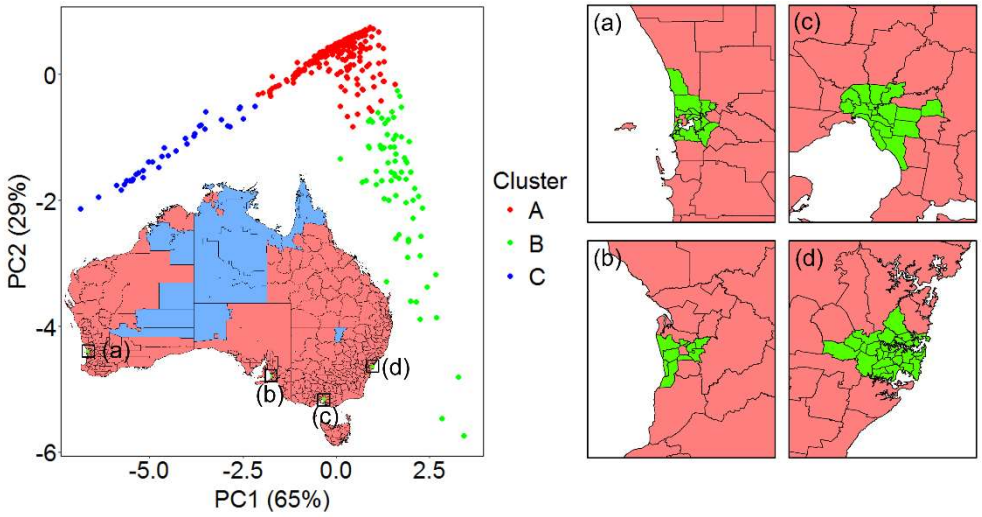
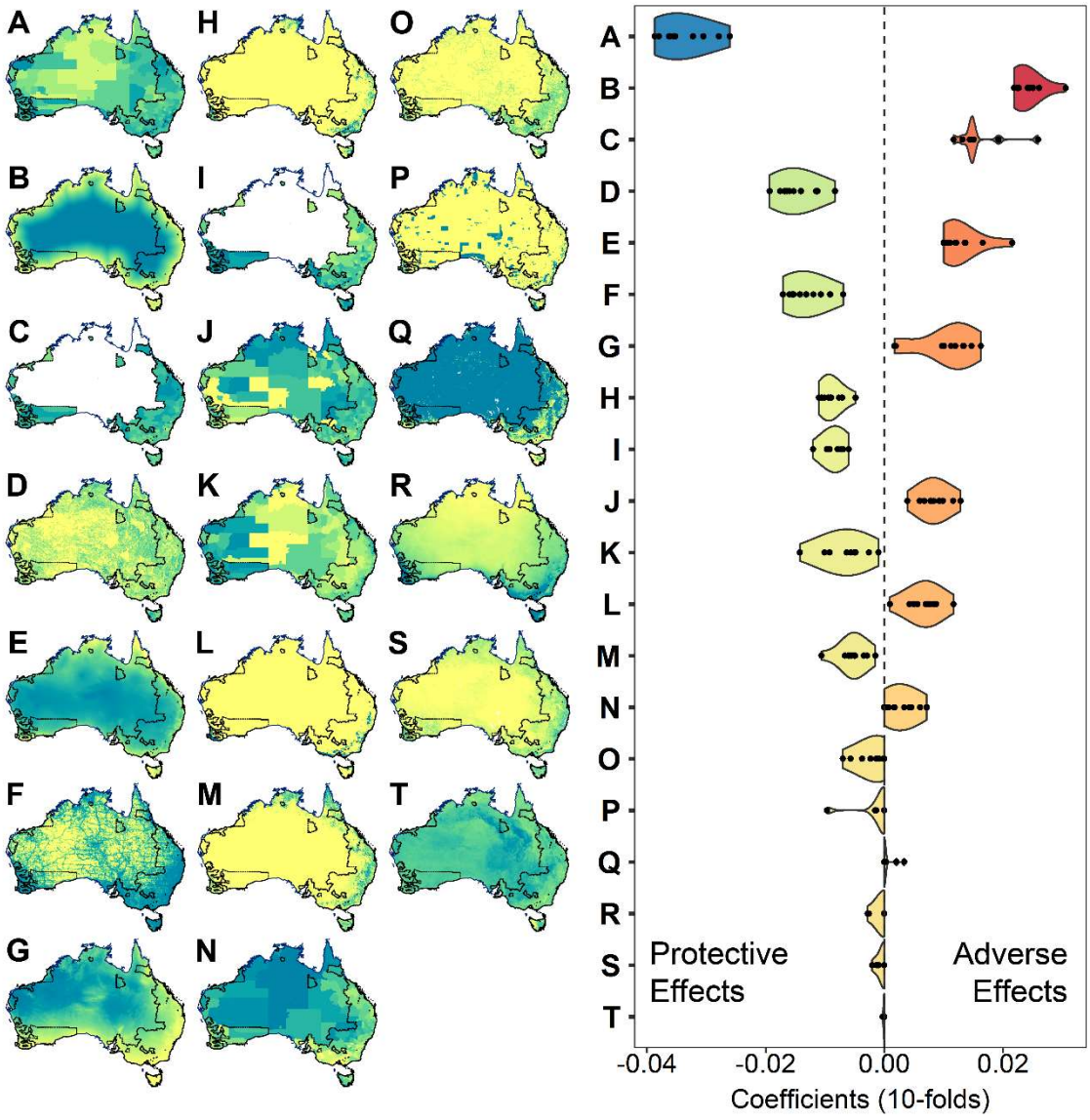


Fig. 1. Socio-geographic clusters of Australian LGAs. Three clusters were identified: the 'Moderate majority' (A, n=451), 'Major cities' (B, n=72), and 'Remote disadvantaged' (C, n=35). Numbers indicate all available LGAs used to determine clusters, whereas a reduced set were available for health response modelling (see Table 1). Inset maps show LGAs associated with (a) Perth, (b) Adelaide, (c) Melbourne, and (d) Sydney.



588

589 **Fig. 2. Important predictors for the 'Moderate majority' socio-geographic cluster.** The
590 right panel shows density and point plots of standardized regression coefficients from 10-fold
591 CV Lasso modelling of log10(respiratory disease public hospital admissions). Shading of
592 density plots indicates negative (blue) to positive (red) coefficients. Variables with negative
593 coefficients associate with decreased hospital admissions. Twenty predictors were identified
594 across the 10-fold Lasso models: (A) Socioeconomic index, (B) Distance to coast, (C)
595 Percent obese persons, (D) Diversity of major vegetation groups, (E) Mean temperature
596 annual range, (F) Species richness (log10), (G) Maximum temperature of warmest month,
597 (H) Proportion of eucalypt forests 10-30 m (logit), (I) Percent overweight persons (logit), (J)
598 Percent smoking during pregnancy, (K) Percent English-speaking immigrants (logit), (L)
599 Proportion of warm wet plains (logit), (M) Proportion of open trees (logit), (N) Percent
600 Aboriginal persons (logit), (O) Diversity of land use, (P) Proportion of nature conservation
601 (logit), (Q) Vegetation fractional cover minimum nonphotosynthetic (logit), (R) Mean
602 precipitation of the coldest quarter, (S) Vegetation fractional cover minimum photosynthetic
603 (logit), (T) Soil cation exchange capacity * erodible fraction (geometric mean). Maps display

all available predictor data for Australia (low values are shown in yellow, high values in blue), however only a portion of local government areas (LGAs) occur in the 'Moderate majority' socio-geographic cluster (Fig. 1). Due to reduced data availability among candidate predictors, a smaller subset of LGAs (n=364, see maps C and I) were used in the modelling.

610 **Table 1. Performance of 10-fold CV Lasso modelling.**

611 Mean performance statistics from explicit 10-fold CV Lasso modelling of respiratory disease
612 public hospital admissions (see Appendix A for further details).

613

CV statistics (mean of 10-fold validation sets)	**Local government area cluster**		
	Moderate majority	**Major cities**	**Remote disadvantaged**
Response transformation	log10	-	log10
No. of LGAs used in modelling (n)	364	62	24
No. of predictors (p) chosen by the Lasso	20	11	10
SD(obs)	0.1479	242.2	0.1824
Root mean square error	0.0999	125.3	0.1463
Mean error (bias)	-0.0003	-2.833	0.0199
Skewness(residuals)	0.1363	-0.1932	-0.0341
Concordance correlation coefficient	0.6795	0.7768	0.4070
R^2	0.5557	0.7698	0.9244

614

615

616

References

ABS, 2015. 2015 National Health Survey: First Results, 2014-15 (Catalogue No. 4364.0.55.001), Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 08/12/2015, Australian Bureau of Statistics, Canberra, Australia.

Access Economics, 2007. The economic impact of allergic disease in Australia: not to be sneezed at. Report by Access Economics Pty Limited for the Australasian Society of Clinical Immunology and Allergy (ASCIA).

AIHW, 2007. Rural, regional and remote health: a study on mortality (2nd edition), Rural health series. Australian Institute of Health and Welfare, Canberra, p. x + 351.

AIHW, Poulos, L.M., Cooper, S.J., Ampon, R., Reddel, H.K., Marks, G.B., 2014. Mortality from asthma and COPD in Australia. Australian Institute of Health and Welfare, Canberra, Australia.

Artis, D., 2008. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nat Rev Immunol* 8, 411-420.

Bardin, P.G., 2004. Vaccination for asthma exacerbations. *Internal Medicine Journal* 34, 358-360.

Belizario, J.E., Napolitano, M., 2015. Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. *Frontiers in Microbiology* 6.

Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P.M., Reith, F., Dennis, P.G., Breed, M.F., Brown, B., Brown, M.V., Brugger, J., Byrne, M., Caddy-Retalic, S., Carmody, B., Coates, D.J., Correa, C., Ferrari, B.C., Gupta, V.V.S.R., Hamonts, K., Haslem, A., Hugenholtz, P., Karan, M., Koval, J., Lowe, A.J., Macdonald, S., McGrath, L., Martin, D., Morgan, M., North, K.I., Paungfoo-Lonhienne, C., Pendall, E., Phillips, L., Pirzl, R., Powell, J.R., Ragan, M.A., Schmidt, S., Seymour, N., Snape, I., Stephen, J.R., Stevens, M., Tinning, M., Williams, K., Yeoh, Y.K., Zammit, C.M., Young, A., 2016. Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. *GigaScience* 5, 21.

Bowers, R.M., McLetchie, S., Knight, R., Fierer, N., 2011. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *ISME J* 5, 601-612.

Branchfield, K., Nantie, L., Verheyden, J.M., Sui, P., Wienhold, M.D., Sun, X., 2016. Pulmonary neuroendocrine cells function as airway sensors to control lung immune response. *Science* 351, 707.

Bringel, F., Couée, I., 2015. Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics. *Frontiers in Microbiology* 6.

Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., Goulding, D., Lawley, T.D., 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* 533, 543-546.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Themaat, E.V.L.v., Schulze-Lefert, P., 2013. Structure and Functions of the Bacterial Microbiota of Plants. *Annual Review of Plant Biology* 64, 807-838.

Cariñanos, P., Casares-Porcel, M., 2011. Urban green zones and related pollen allergy: A review. Some guidelines for designing spaces with low allergy impact. *Landscape and Urban Planning* 101, 205-214.

Chen, L., Verrall, K., Tong, S., 2006. Air particulate pollution due to bushfires and respiratory hospital admissions in Brisbane, Australia. *International Journal of Environmental Health Research* 16, 181-191.

Clemente, Jose C., Ursell, Luke K., Parfrey, Laura W., Knight, R., 2012. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell* 148, 1258-1270.

Craig, J.M., Logan, A.C., Prescott, S.L., 2016. Natural environments, nature relatedness and the ecological theater: connecting satellites and sequencing to shinrin-yoku. *Journal of Physiological Anthropology* 35, 1-10.

Davies, J.M., Beggs, P.J., Medek, D.E., Newnham, R.M., Erbas, B., Thibaudon, M., Katelaris, C.H., Haberle, S.G., Newbigin, E.J., Huete, A.R., 2015. Trans-disciplinary research in synthesis of grass pollen aerobiology and its importance for respiratory health in Australasia. *Science of The Total Environment* 534, 85-96.

de Vries, S., Verheij, R.A., Groenewegen, P.P., Spreeuwenberg, P., 2003. Natural Environments—Healthy Environments? An Exploratory Analysis of the Relationship between Greenspace and Health. *Environment and Planning A* 35, 1717-1731.

Després, V., Huffman, J.A., Burrows, S., Hoose, C., Safatov, A., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M., Pöschl, U., Jaenicke, R., 2012. Primary biological aerosol particles in the atmosphere: a review. *Tellus B: Chemical and Physical Meteorology* 64, 15598.

Dickson, R.P., Erb-Downward, J.R., Huffnagle, G.B., 2013. The role of the bacterial microbiome in lung disease. *Expert Review of Respiratory Medicine* 7, 245-257.

Douwes, J., Travier, N., Huang, K., Cheng, S., McKenzie, J., Le Gros, G., Von Mutius, E., Pearce, N., 2007. Lifelong farm exposure may strongly reduce the risk of asthma in adults. *Allergy* 62, 1158-1165.

Ege , M.J., Mayer , M., Normand , A.-C., Genuneit , J., Cookson , W.O.C.M., Braun-Fahrländer , C., Heederik , D., Piarroux , R., von Mutius , E., 2011. Exposure to Environmental Microorganisms and Childhood Asthma. *New England Journal of Medicine* 364, 701-709.

Elliot, P., Wakefield, J., Best, N., Briggs, D., 2000. Spatial epidemiology: methods and applications. Oxford University Press, New York, p. 475.

Elliott, K.C., Cheruvelil, K.S., Montgomery, G.M., Soranno, P.A., 2016. Conceptions of Good Science in Our Data-Rich World. *BioScience* 66, 880-889.

Flegal, K.M., Kalantar-Zadeh, K., 2013. Overweight, Mortality and Survival. *Obesity* 21, 1744-1745.

695 Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear
696 Models via Coordinate Descent. *Journal of Statistical Software* 33, 1-22.

697 Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. *The Annals of*
698 *Applied Statistics* 2, 916–954.

699 Fujimura, K.E., Demoor, T., Rauch, M., Faruqi, A.A., Jang, S., Johnson, C.C., Boushey,
700 H.A., Zoratti, E., Ownby, D., Lukacs, N.W., Lynch, S.V., 2014. House dust exposure
701 mediates gut microbiome Lactobacillus enrichment and airway immune defense against
702 allergens and virus infection. *PNAS* 111, 805-810.

703 Fujimura, K.E., Lynch, S.V., 2015. Microbiota in Allergy and Asthma and the Emerging
704 Relationship with the Gut Microbiome. *Cell Host & Microbe* 17, 592-602.

705 Gaisberger, M., Šanović, R., Dobias, H., Kolarž, P., Moder, A., Thalhamer, J., Selimović, A.,
706 Huttegger, I., Ritter, M., Hartl, A., 2012. Effects of Ionized Waterfall Aerosol on Pediatric
707 Allergic Asthma. *Journal of Asthma* 49, 830-838.

708 Geoscience Australia, 2014. Dynamic Land Cover Dataset V1.0, 27 May 2014.

709 Gilbert, J.A., Jansson, J.K., Knight, R., 2014. The Earth Microbiome project: successes and
710 aspirations. *BMC Biology* 12, 69.

711 Gladwell, V.F., Brown, D.K., Wood, C., Sandercock, G.R., Barton, J.L., 2013. The great
712 outdoors: how a green exercise environment can benefit all. *Extreme Physiology & Medicine*
713 2:3.

714 Groombridge, B. (ed), 1992. Global biodiversity: Status of the earth's living resources. A
715 report compiled by the World Conservation Monitoring Centre. Chapman and Hall, London.

716 Gubhaju, L., McNamara, B.J., Banks, E., Joshy, G., Raphael, B., Williamson, A., Eades, S.J.,
717 2013. The overall health and risk factor profile of Australian Aboriginal and Torres Strait
718 Islander participants from the 45 and up study. *BMC Public Health* 13, 1-14.

719 Haahtela, T., Holgate, S., Pawankar, R., Akdis, C.A., Benjaponpitak, S., Caraballo, L., 2013.
720 The biodiversity hypothesis and allergic disease: World Allergy Organization position
721 statement. *World Allergy Organ J* 6.

722 Hanski, I., von Hertzen, L., Fyhrquist, N., Koskinen, K., Torppa, K., Laatikainen, T.,
723 Karisola, P., Auvinen, P., Paulin, L., Mäkelä, M.J., Vartiainen, E., Kosunen, T.U., Alenius,
724 H., Haahtela, T., 2012. Environmental biodiversity, human microbiota, and allergy are
725 interrelated *PNAS* 109, 8334-8339.

726 Ichinohe, T., Pang, I.K., Kumamoto, Y., Peaper, D.R., Ho, J.H., Murray, T.S., Iwasaki, A.,
727 2011. Microbiota regulates immune defense against respiratory tract influenza A virus
728 infection. *PNAS* 108, 5354-5359.

729 Kellett, F., Robert, N.M., 2011. Nebulised 7% hypertonic saline improves lung function and
730 quality of life in bronchiectasis. *Respiratory Medicine* 105, 1831-1835.

731 Keniger, L., Gaston, K., Irvine, K., Fuller, R., 2013. What are the Benefits of Interacting with
732 Nature? *International Journal of Environmental Research and Public Health* 10, 913.

Kim, Y.-G., Udayanga, Kankanam Gamage S., Totsuka, N., Weinberg, Jason B., Núñez, G., Shibuya, A., 2014. Gut Dysbiosis Promotes M2 Macrophage Polarization and Allergic Airway Inflammation via Fungi-Induced PGE2. *Cell Host & Microbe* 15, 95-102.

Leng, C., Lin, Y., Wahba, G., 2006. A note on the LASSO and related procedures in model selection. *Statistica Sinica* 16, 1273-1284.

Liddicoat, C., Waycott, M., Weinstein, P., 2016. Environmental Change and Human Health: Can Environmental Proxies Inform the Biodiversity Hypothesis for Protective Microbial–Human Contact? *BioScience* 66, 1023-1034.

Liora, N., Markakis, K., Poupkou, A., Giannaros, T.M., Melas, D., 2015. The natural emissions model (NEMO): Description, application and model evaluation. *Atmospheric Environment* 122, 493-504.

Maas, J., Verheij, R.A., Groenewegen, P.P., de Vries, S., Spreeuwenberg, P., 2006. Green space, urbanity, and health: how strong is the relation? *Journal of Epidemiology and Community Health* 60, 587-592.

Mansiaux, Y., Carrat, F., 2014. Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. *BMC Medical Research Methodology* 14, 99.

Mhuireach, G., Johnson, B.R., Altrichter, A.E., Ladau, J., Meadow, J.F., Pollard, K.S., Green, J.L., 2016. Urban greenness influences airborne bacterial community composition. *Science of The Total Environment* 571, 680-687.

Michelozzi, P., Accetta, G., De Sario, M., D'Ippoliti, D., Marino, C., Baccini, M., Biggeri, A., Anderson, H.R., Katsouyanni, K., Ballester, F., Bisanti, L., Cadum, E., Forsberg, B., Forastiere, F., Goodman, P.G., Hojs, A., Kirchmayer, U., Medina, S., Paldy, A., Schindler, C., Sunyer, J., Perucci, C.A., 2009. High Temperature and Hospitalizations for Cardiovascular and Respiratory Causes in 12 European Cities. *American Journal of Respiratory and Critical Care Medicine* 179, 383-389.

Mitchell, R., Popham, F., 2008. Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet* 372, 1655-1660.

Molloy, M.J., Bouladoux, N., Belkaid, Y., 2012. Intestinal microbiota: Shaping local and systemic immune responses. *Seminars in Immunology* 24, 58-66.

Myers, S.S., Gaffikin, L., Golden, C.D., Ostfeld, R.S., H. Redford, K., H. Ricketts, T., Turner, W.R., Osofsky, S.A., 2013. Human health impacts of ecosystem alteration. *PNAS* 110, 18753-18760.

Noval Rivas, M., Crother, T.R., Arditi, M., 2016. The microbiome in asthma. *Current Opinion in Pediatrics* 28, 764–771.

Parenteau, M.-P., Sawada, M.C., 2011. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO2 and respiratory health. *International Journal of Health Geographics* 10, 58.

772 Parker, D., Prince, A., 2011. Innate Immunity in the Respiratory Epithelium. *American*
773 *Journal of Respiratory Cell and Molecular Biology* 45, 189-201.

774 PHIDU, 2015. Social Health Atlas of Australia: Data by Local Government Area, March
775 2015 release, Public Health Information Development Unit. Torrens University Australia,
776 Adelaide.

777 PHIDU, 2016. Social Health Atlas of Australia: Data by Local Government Area, May 2016
778 release, Public Health Information Development Unit. Torrens University Australia,
779 Adelaide.

780 Polymenakou, P.N., 2012. Atmosphere: A Source of Pathogenic or Beneficial Microbes?
781 *Atmosphere* 3, 87-102.

782 Prevots, D.R., Marras, T.K., 2015. Epidemiology of human pulmonary infection with non-
783 tuberculous mycobacteria: A review. *Clinics in Chest Medicine* 36, 13-34.

784 Rook, G., 2013. Regulation of the immune system by biodiversity from the natural
785 environment: an ecosystem service essential to health. *PNAS* 110, 18360-18367.

786 Rottem, M., Geller-Bernstein, C., Shoenfeld, Y., 2015. Atopy and Asthma in Migrants: The
787 Function of Parasites. *International Archives of Allergy and Immunology* 167, 41-46.

788 Ruokolainen, L., von Hertzen, L., Fyhrquist, N., Laatikainen, T., Lehtomäki, J., Auvinen, P.,
789 Karvonen, A.M., Hyvärinen, A., Tillmann, V., Niemelä, O., Knip, M., Haahtela, T.,
790 Pekkanen, J., Hanski, I., 2015. Green areas around homes reduce atopic sensitization in
791 children. *Allergy* 70, 195-202.

792 Sandifer, P.A., Sutton-Grier, A.E., Ward, B.P., 2015. Exploring connections among nature,
793 biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance
794 health and biodiversity conservation. *Ecosystem Services* 12, 1-15.

795 Seedorf, H., Griffin, Nicholas W., Ridaura, Vanessa K., Reyes, A., Cheng, J., Rey,
796 Federico E., Smith, Michelle I., Simon, Gabriel M., Scheffrahn, Rudolf H., Woebken, D.,
797 Spormann, Alfred M., Van Treuren, W., Ursell, Luke K., Pirrung, M., Robbins-Pianka, A.,
798 Cantarel, Brandi L., Lombard, V., Henrissat, B., Knight, R., Gordon, Jeffrey I., 2014.
799 Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* 159, 253-266.

800 Simpson, R., Williams, G., Petroeschevsky, A., Best, T., Morgan, G., Denison, L., Hinwood,
801 A., Neville, G., Neller, A., 2005. The short-term effects of air pollution on daily mortality in
802 four Australian cities. *Australian and New Zealand Journal of Public Health* 29, 205-212.

803 Sonnenburg, E.D., Smits, S.A., Tikhonov, M., Higginbottom, S.K., Wingreen, N.S.,
804 Sonnenburg, J.L., 2016. Diet-induced extinctions in the gut microbiota compound over
805 generations. *Nature* 529, 212-215.

806 State of the Environment 2011 Committee, 2011. Australia state of the environment 2011.
807 Independent report to the Australian Government Minister for Sustainability, Environment,
808 Water, Population and Communities. DSEWPaC, Canberra.

809 Stein, M.M., Hrusch, C.L., Gozdz, J., Igartua, C., Pivniouk, V., Murray, S.E., Ledford, J.G.,
810 Marques dos Santos, M., Anderson, R.L., Metwali, N., Neilson, J.W., Maier, R.M., Gilbert,
811 J.A., Holbreich, M., Thorne, P.S., Martinez, F.D., von Mutius, E., Vercelli, D., Ober, C.,

812 Sperling, A.I., 2016. Innate Immunity and Asthma Risk in Amish and Hutterite Farm
813 Children. *New England Journal of Medicine* 375, 411-421.

814 Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal*
815 *Statistical Society. Series B (Methodological)* 58, 267-288.

816 von Hertzen, L., Haahtela, T., 2006. Disconnection of man and the soil: Reason for the
817 asthma and atopy epidemic? *Journal of Allergy and Clinical Immunology* 117, 334-344.

818 von Hertzen, L., Hanski, I., Haahtela, T., 2011. Natural immunity: Biodiversity loss and
819 inflammatory diseases are two global megatrends that might be related. *EMBO reports* 12,
820 1089-1093.

821 Voreades, N., Kozil, A., Weir, T., 2014. Diet and the development of the human intestinal
822 microbiome. *Frontiers in Microbiology* 5:494.

823 Whitsett, J.A., Alenghat, T., 2014. Respiratory epithelial cells orchestrate pulmonary innate
824 immunity. *Nature immunology* 16, 27-35.

825 WHO, SCBD, 2015. Connecting global priorities: biodiversity and human health. A state of
826 knowledge review. World Health Organisation, Geneva, Switzerland.

827 Wopereis, H., Oozeer, R., Knipping, K., Belzer, C., Knol, J., 2014. The first thousand days –
828 intestinal microbiology of early life: establishing a symbiosis. *Pediatric Allergy and*
829 *Immunology* 25, 428-438.

830 Zammit, C., Liddicoat, H., Moonsie, I., Makker, H., 2010. Obesity and respiratory diseases.
831 *International Journal of General Medicine* 3, 335-343.

832

833

Supplementary Material for

Landscape biodiversity correlates with respiratory health in Australia

Craig Liddicoat[a,b,*], Peng Bi[c], Michelle Waycott[a,b], John Glover[d], Andrew J. Lowe[a], Philip Weinstein[a]

[a]School of Biological Sciences and The Environment Institute, The University of Adelaide, North Terrace, Adelaide SA 5005, Australia.

[b]Department of Environment, Water and Natural Resources, GPO Box 1047, Adelaide SA 5001, Australia.

[c]School of Public Health, The University of Adelaide, North Terrace, Adelaide SA 5005, Australia.

[d]Public Health Information Development Unit, Torrens University Australia, Level 1, 200 Victoria Square, Adelaide SA 5000, Australia.

*Author for correspondence: Craig Liddicoat. e-mail: craig.liddicoat@adelaide.edu.au

Contents:

Methods (Supplementary Information)

Environmental covariate preparation

We collated and prepared a wide array of environmental covariates from various sources (Table S3). A standard grid system was adopted matching the approx. 250 m cell size of Geoscience Australia dynamic land cover mapping data [1]. Where necessary, numeric layers were bilinearly resampled and categorical layers were resampled using the nearest neighbor method, in order to match the common grid system.

Focal neighborhood calculations using a 3 km radius were then undertaken for all of the Australia-wide map layers, such that each grid cell subsequently contained a summary value characterizing the surrounding 3 km-radius area. For numeric layers, each grid cell was recalculated to contain the average value of the surrounding area. Categorical layers were converted into numeric layers via (a) iteratively evaluating the proportion of each class within the 3 km-radius area, and (b) calculating the diversity of classes using the Shannon diversity index (H):

$$H = -\sum_{i=1}^{k} p_i \ln(p_i) \qquad \text{(S1.1)}$$

where p_i is the proportion in class k, and k is the number of classes found in each respective 3km-radius area.

A derived layer for the erodible fraction (*EF*) of soils was calculated using the formula proposed by Fryrear *et al.* [2]:

$$EF = [\ 29.09 + 0.31 * sand + 0.17 * silt + 0.33 * \left(\frac{sand}{clay}\right) - 4.66 * OC - 0.95 * CaCO_3\ ]/100 \qquad \text{(S1.2)}$$

where: *sand* = sand content (%), *silt* = silt content (%); *OC* = organic carbon content (%), $CaCO_3$ = calcium carbonate content (%); and sand, silt and *OC* content layers were sourced from Viscarra Rossel *et al.* [3], while the $CaCO_3$ content layer was sourced from Wilford *et al.* [4].

The majority of spatial data preparation and analysis was performed using the *R* software environment [5], in particular using tools adapted from the *R raster* package [6]. Australia-wide focal neighborhood calculations for class proportions and Shannon diversity indices were implemented using *R* software on high-performance computing facilities provided by eRSA (www.ersa.edu.au).

Environmental variables were finally summarized by averaging values to match Local Government Area (LGA) boundaries corresponding to the available public health and contextual data, using the zonal statistics tool in ESRI ArcGIS 10.2 [7].

Data cleaning due to missing data

The Social Health Atlas of Australia has missing data for some LGAs across the response and candidate predictor variables. To consider the widest possible range of potential explanatory variables, in the 'Moderate majority' and 'Major cities' clusters we first excluded all rows (LGAs) with missing data. This reduced the number of LGAs used (from 451 to 364 for the 'Moderate majority'; and from 72 to 62 for 'Major cities') in the subsequent modelling and analysis.

In the 'Remote disadvantaged' cluster, we first excluded four variables with a high frequency of missing data (Percent smokers, Percent high alcohol consumption, Percent overweight persons, Percent obese persons) and then excluded any remaining rows with missing data. This process reduced the number of LGAs available for analysis from 35 to 24.

Automated screening of candidate predictors

Because the Lasso tool is effectively a machine-learning adaptation of multiple linear regression, we considered the need to screen and potentially transform candidate predictors to help improve linear relationships while balancing the need for interpretability (of predictors) and transparency in any transformation process. To achieve this in a transparent, objective fashion, we developed an automated screening algorithm (using *R* script) to scrutinise and where necessary transform or eliminate candidate predictors prior to input to the Lasso. The algorithm was run separately for each LGA cluster and comprised the following steps:

Step 1: Variables were excluded if there were less than 50% of the total LGAs with non-zero observations from which to infer a relationship. This was designed to eliminate predictors that were present in the Australia-wide data but were considered non-representative for the particular socio-geographic cluster under consideration.

Step 2: The following transformations were calculated for each candidate predictor for later comparison against each other and the original (untransformed) data:

- Logit (log odds) transformations were the only alternative considered for proportion and percentage-based data, with lower and upper bounds (i.e. 0, 100%) remapped to the 2.5th and 97.5th percentile respectively.
- Square root (sqrt), which was disqualified if negative values were present.
- Log10, which was disqualified if negative or zero values were present.
- Square
- Cube root, which was only considered where negative values were present.

Step 3: Pragmatic threshold acceptance criteria were set for the original and transformed variables to be included in the modelling. Any candidate predictors not meeting these criteria were excluded:

- absolute value of skewness <= 1.5
- kurtosis <= 4

- coefficient of determination (R^2) between the candidate predictor and response variable >= 0.025 (i.e. at least 2.5% stand-alone explanatory value).

Step 4: Representatives for each variable were then chosen. If only one option remained (whether original or transformed), that was chosen. Then, for proportion and percentage-based variables, the most normally distributed data (comparing only original and logit transforms) were chosen, as judged by the maximum Shapiro-Wilk normality test p-value. For non-proportion-based variables, if the original form passed the acceptance criteria then that was chosen as a priority for ease of interpretation. If the original was not eligible, then the most normally-distributed data from the remaining transformations was chosen (again, as judged by the maximum Shapiro-Wilk normality test p-value). For a number of variables, no options passed the acceptance criteria, and they were eliminated from the modelling.

All of the above analyses, and subsequent modelling, were performed on candidate predictor variables that had been subject to 95% Winsorization (based on data within each cluster), which was designed to eliminate the influence of extreme and potentially outlying values [8]. This involved all extreme low values below the 2.5th percentile being replaced by the 2.5th percentile, and all extreme high values above the 97.5th percentile being replaced by the 97.5th percentile.

Predictor data were then centered and scaled (based on data within each cluster) prior to the 10-fold Lasso modelling.

Statistical analysis

The performance of the 10-fold Lasso modelling was assessed using mean cross-validation statistics for: root mean-square error, mean error (or bias), skewness of model residuals, coefficient of determination (R^2), and concordance correlation coefficient [9]. That is, health response observations from LGAs in each respective validation set (not used to develop the respective k-fold Lasso model) were compared against respective predictions based on the validation set predictor data. These results were then averaged over the 10 validation sets. This process was repeated for each socio-geographic cluster (see Table 1 of main article). The R^2 values were computed from the square of the Pearson correlation coefficient and indicate how well predictions and observations adhere to a straight line (least-squares linear fit). The concordance correlation coefficient indicates how well predictions and observations adhere to a 1:1 relationship, with values closer to 1 indicating greater agreement. The skewness of residuals indicates how balanced the model is in terms of under- and over-predictions. Table 1 also reports the standard deviation of observations (SD (obs)). Performance measures were calculated using the same form of health response data as used in modelling, i.e. using log10-transformed response data for the 'Moderate majority' and 'Remote disadvantaged', and untransformed data for the 'Major cities'. Validation set predictions and observations are plotted in Fig. S6.

To examine the significance of Diversity of major vegetation groups within quartiles of Socioeconomic index (using 'Moderate majority' data only, Fig. S7) we constructed standard multiple linear regression models for the health response (log10(respiratory disease public hospital admissions)) using only Socioeconomic index and Diversity of major

vegetation groups as explanatory variables. Variable significance was assessed using p-values for the regression coefficients from the respective multiple linear regression models. Throughout this study, where standard linear regression models are applied, these use base R software [5]. Testing of Lasso-identified predictor significance (i.e. selection-adjusted p-values, Tables S6-S7) use the R SelectiveInference package [10].

Extended analysis: examining variable significance

Based on the 10-fold Lasso modelling results, and using the 'Moderate majority' data only, we undertook two further analyses to better understand variable significance. Firstly, we examined the influence of Diversity of major vegetation groups, while controlling for Socioeconomic influence (Fig. S7). Within quartiles of Socioeconomic index, we considered only Socioeconomic index and Diversity of major vegetation groups as explanatory variables in separate multiple linear regression models for respiratory health. In every quartile, Diversity of major vegetation groups was identified as significant predictor. However, only in the wealthiest quartile was Socioeconomic index also identified as a significant predictor.

Secondly, we wished to understand whether predictors identified by the Lasso could be viewed as significant. We noted that the assessment of significance using contemporary machine-learning tools is a developing field of statistical science, and new tools for investigating selective inference with the Lasso provide more conservative p-values than might be expected from traditional methods [11]. To examine the significance of results from the Lasso we re-analysed all 'Moderate majority' data, this time considering a two-part random split (each with n=182 LGAs). (The reason for splitting data is that the Lasso tool is able to provide important insights into a particular dataset, and can cherry-pick variables with the most explanatory value for that dataset. From a purist statistical science viewpoint, it is therefore not appropriate to reuse the same dataset and the cherry-picked (Lasso-identified) variables in a traditional multiple linear regression model, as this will not provide a fair and independent assessment of variable significance.)

On data-split one we ran a single iteration of the Lasso (Fig. S8), and determined the selection-adjusted p-values [10] for coefficients of identified non-zero predictors. Ordering these Lasso-identified predictors by the absolute size of standardized regression coefficients, we then inputted respective predictor data from data-split two into standard multiple linear regression software. This generated an independently-generated and contrasting set of regression coefficients and p-values (Table S6). For completeness, we also reversed these operations: running the Lasso on data-split two (Fig. S9), and then entering the corresponding identified predictor data from data-split one into a standard multiple linear regression (with results in Table S7). For this extended Lasso analysis we used the default internal 10-fold cross-validation setting when calling cv.glmnet to identify the optimal penalisation parameter (using the parsimonious option, s = 'lambda.1se') and corresponding Lasso regression coefficients.

Results show some consistency with similar themes of predictors identified in comparison to the 10-fold modelling. However, different results were obtained because of differences in input data. P-values provided by the selective inference software were

conservative, however many of the predictors ranked highly by the Lasso (when ordered by absolute size of standardized regression coefficients) also showed up as significant predictors in the complementary data-split (i.e. using data kept separate from the Lasso variable-selection process).

Associated data

The following data are available for download from the article web page on ScienceDirect:

- Database S1 – All data in raw form (558 local government areas x 192 candidate explanatory variables)
- Example R scripts for:
    - Continent-wide 3 km-radius focal neighbourhood calculation - Shannon diversity index
    - Continent-wide 3 km-radius focal neighbourhood calculation - class proportion #1
    - Modelling workflow for Australia-wide 10-fold Lasso modelling of respiratory health by local government areas

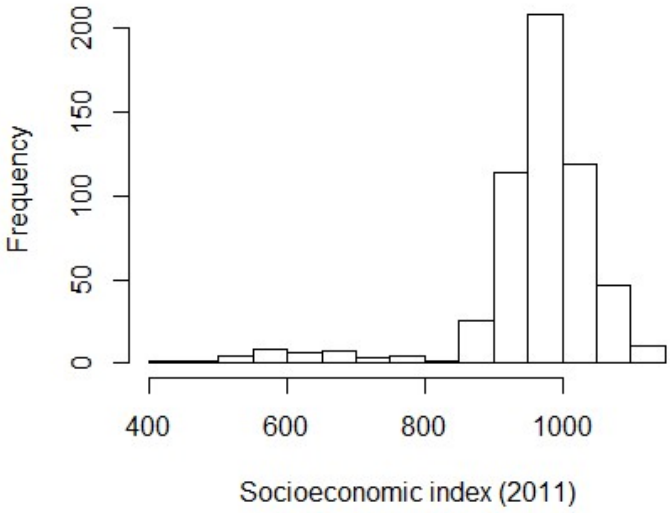


Fig. S1. Histogram of Australia-wide Socioeconomic index

Frequency distribution of Socioeconomic index (termed 'Index of relative socioeconomic disadvantage' in the Social Health Atlas of Australia [12, 13]) for n=558 local government areas across Australia. The left tail is suggestive of a second minor mode centered on communities of low socioeconomic status. Note that low values of Socioeconomic index correspond to greater disadvantage or deprivation, while high values correspond to higher socioeconomic status.

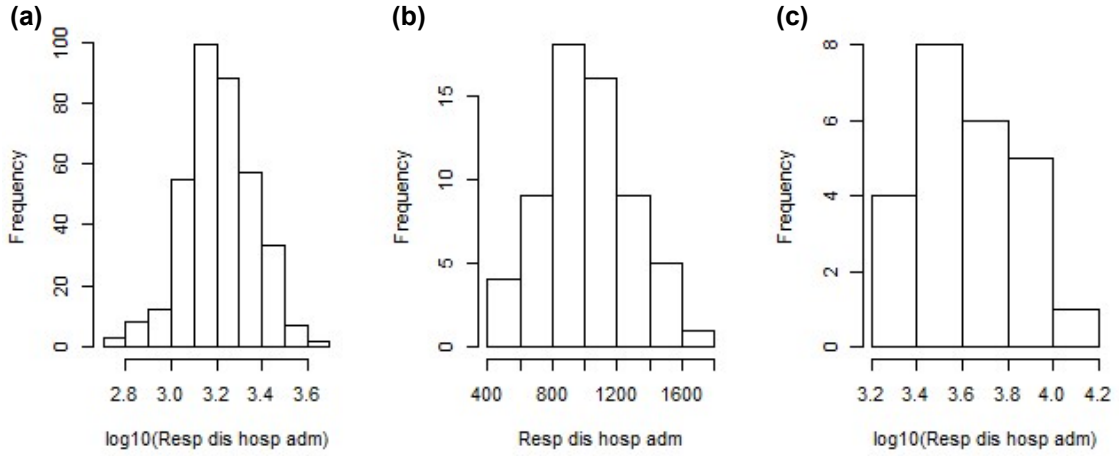


Fig. S2. Histograms of respiratory disease response variables

Frequency distributions of respiratory disease public hospital admissions (health response variable) for local government area clusters: (a) 'Moderate majority' [log10-transformed, n=364, mean ± sd: 3.210 ± 0.150], (b) 'Major cities' [n=62, mean ± sd: 1022 ± 267], and (c) 'Remote disadvantaged' [log10-transformed, n=24, mean ± sd: 3.612 ± 0.213]. Raw units are age standardized rate per 100,000.

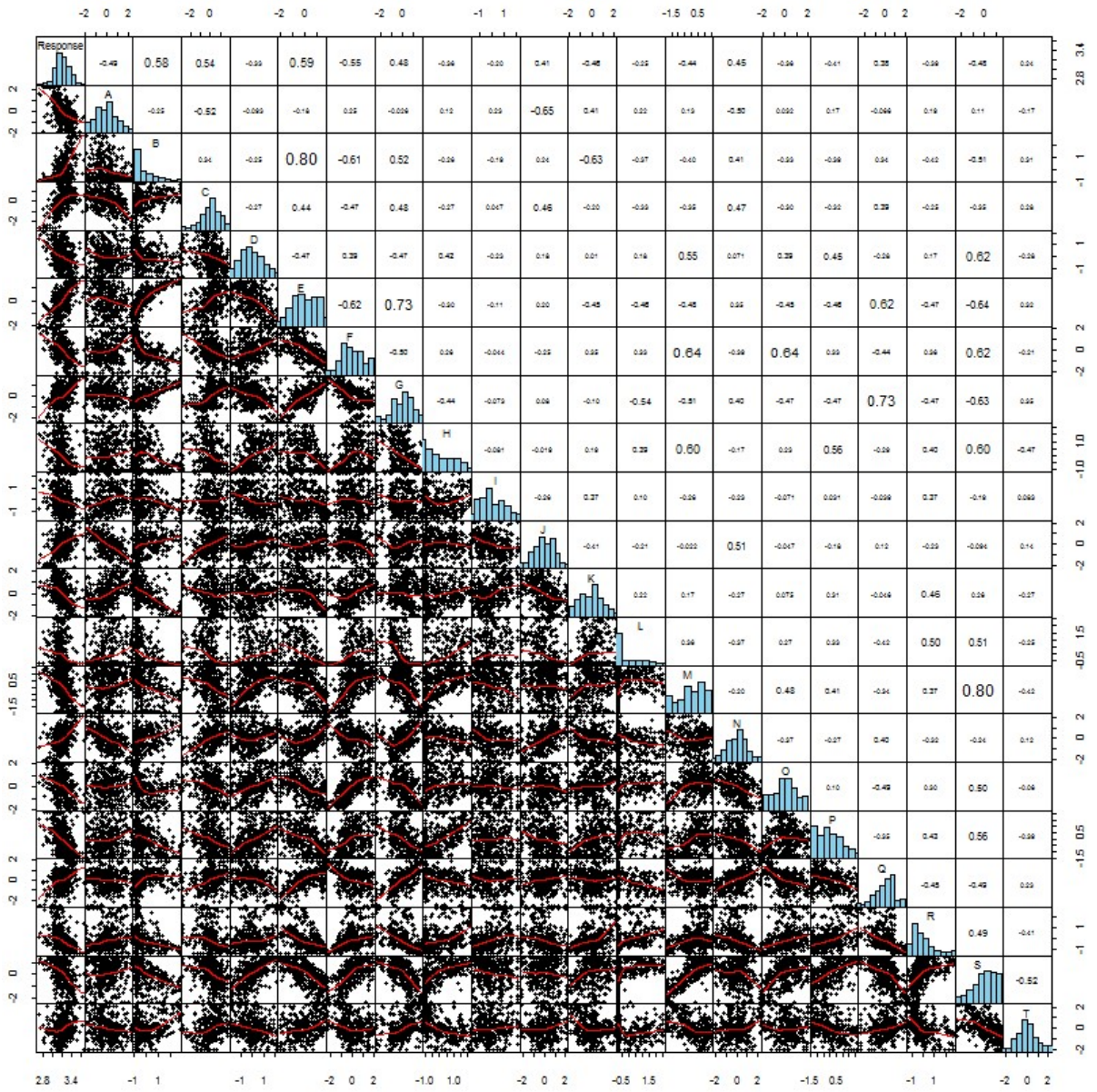


Fig. S3. Scatterplot matrix with correlation coefficients for response and predictor variables identified from 10-fold cross-validation Lasso modelling of the 'Moderate majority' cluster.

Variables labelled on the diagonal are 'Response', i.e. log10(respiratory disease public hospital admissions)', (A) Socioeconomic index, (B) Distance to coast, (C) Percent obese persons, (D) Diversity of major vegetation groups, (E) Mean temperature annual range, (F) Species richness (log10), (G) Maximum temperature of warmest month, (H) Proportion of eucalypt forests 10-30m (logit), (I) Percent overweight persons (logit), (J) Percent smoking during pregnancy, (K) Percent English-speaking immigrants (logit), (L) Proportion of warm wet plains (logit), (M) Proportion of open trees (logit), (N) Percent Aboriginal persons (logit), (O) Diversity of land use, (P) Proportion of nature conservation (logit), (Q) Vegetation fractional cover minimum nonphotosynthetic (logit), (R) Mean precipitation of the coldest quarter, (S) Vegetation fractional cover minimum photosynthetic (logit), (T) Soil cation exchange capacity * erodible fraction (geometric mean). Panels below the diagonal contain scatterplots with locally weighted scatterplot smoothing ('lowess') curves; and above the diagonal contain the calculated Pearson correlation coefficient (r). Variable histograms are included on the diagonal.

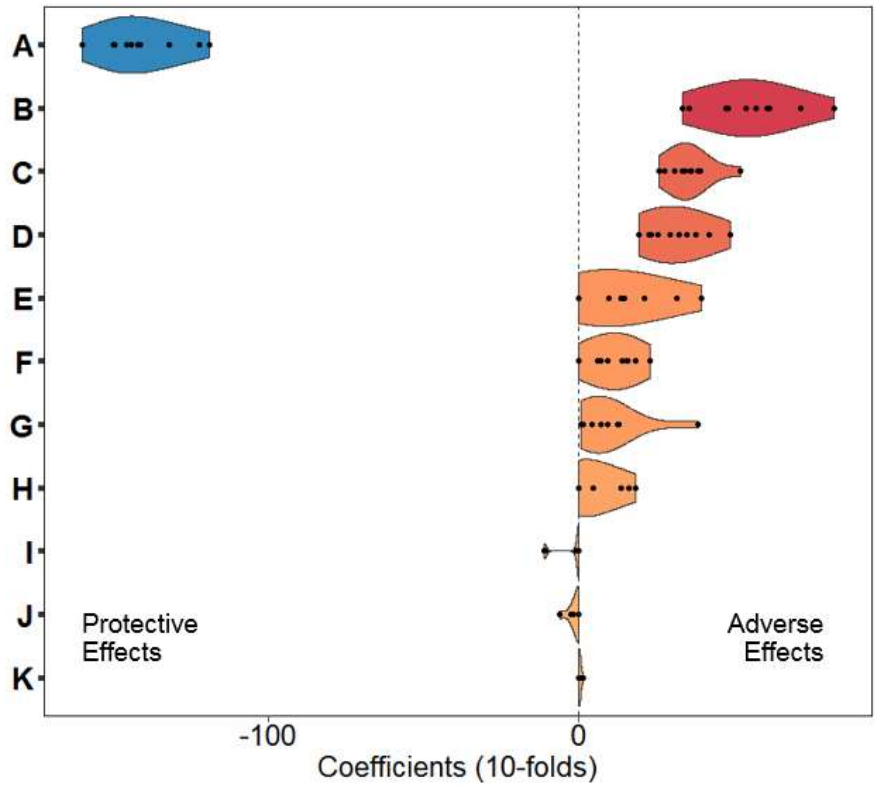


Fig. S4. Important predictors for the 'Major cities' cluster (n=62).

Density and point plots of standardized regression coefficients from 10-fold cross-validation Lasso modelling of respiratory disease public hospital admissions. Eleven predictors were identified across the 10-fold Lasso models: (A) Socioeconomic index, (B) Mean temperature of the wettest quarter, (C) Proportion of rainfed pasture (logit), (D) Percent Aboriginal persons (logit), (E) Percent of smokers, (F) Percent obese persons, (G) Percent smoking during pregnancy, (H) Vegetation fraction of photosynthetically-active radiation maximum (logit), (I) Total population, (J) Percent overseas-born from non-English-speaking countries resident less than 5 years (logit), (K) Proportion of tussock grass open (logit).

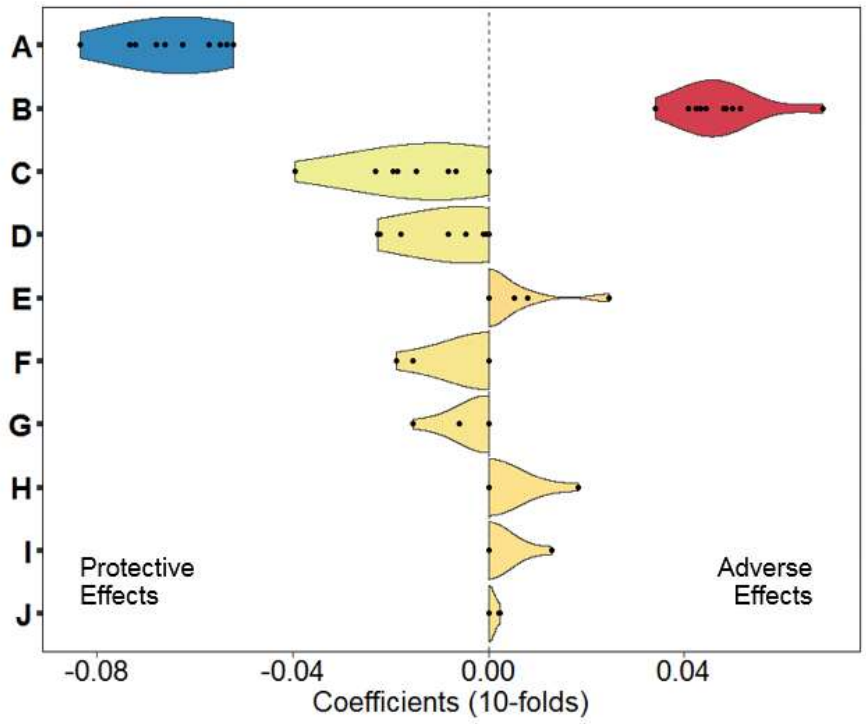


Fig. S5. Important predictors for the 'Remote disadvantaged' cluster (n=24).

Density and point plots of standardized regression coefficients from 10-fold cross-validation Lasso modelling of respiratory disease public hospital admissions. Ten predictors were identified across the 10-fold Lasso models: (A) Proportion of wetlands (logit), (B) Mean temperature of the wettest quarter, (C) Proportion of acacia forests and woodlands (logit), (D) Proportion of swampy grasses and sedges (logit), (E) Proportion of hummock grass sparse (logit), (F) Air temperature annual mean isothermality, (G) Proportion of eucalypt woodlands (logit), (H) Vegetation fractional cover standard deviation – bare soil, (I) Proportion of scattered trees (logit), (J) Vegetation fractional cover maximum – bare soil (logit).

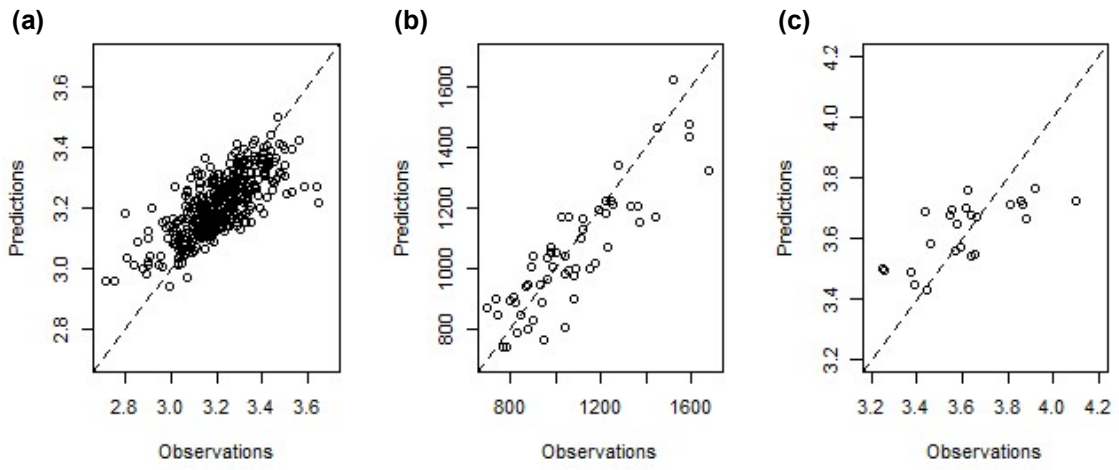
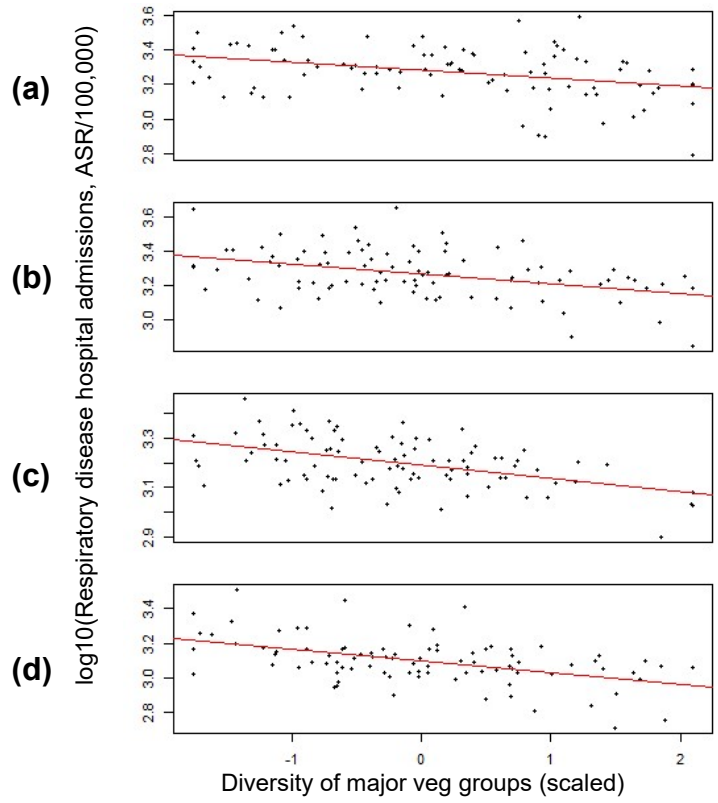


Fig. S6. Cross-validation plots

10-fold cross-validation predictions versus observations for Lasso penalized regression modelling of respiratory disease public hospital admissions, for local government area clusters: (a) ‘Moderate majority’, (b) ‘Major cities’, and (c) ‘Remote disadvantaged’. Note: response variables for clusters (a) and (c) are log10-transformed.



	p-value for regression coefficient		Correlation (r)
	Socio-economic index	Div. of major veg. groups	Div. of major veg. groups & Socioeconomic index
(a)	0.3728 -	0.0005 ***	-0.2037
(b)	0.0437 *	0.0001 ***	0.0746
(c)	0.0665 .	9e-07 ***	-0.1350
(d)	1e-05 ***	2e-06 ***	0.1259

Fig. S7. Relationship between Diversity of major vegetation groups, Socioeconomic index, and respiratory health response for the 'Moderate majority' cluster—within quartiles of Socioeconomic index.

Plots show data for local government areas split by: (a) 0-25th, (b) 25-50th, (c) 50-75th, and (d) 75-100th percentiles (i.e. quartiles) of Socioeconomic index. For each socioeconomic quartile (n=91), the corresponding table row on the right contains p-values for regression coefficients from a multiple linear regression model for the health response—log10(respiratory disease public hospital admissions)—with only Socioeconomic index and Diversity of major vegetation groups as explanatory variables. Significance codes are: 0–0.001: '***', 0.001–0.01: '**', 0.01–0.05: '*', 0.05– 0.1: '.' The third column in the table shows the respective Pearson correlation coefficient (r) calculated between Diversity of major vegetation groups and Socioeconomic index (indicating they have low levels of correlation).

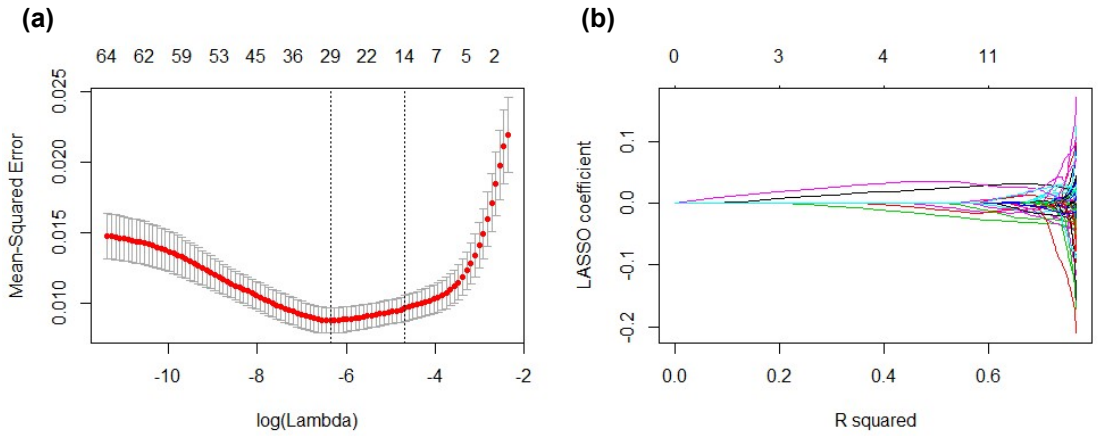


Fig. S8. Alternate Lasso model parameters for the 'Moderate majority' cluster – based on *data-split <u>one</u>* training data.

Panel (a) shows mean square error on the y-axis, the log of the penalisation parameter is shown on the lower x-axis, and the increasing number of non-zero predictors introduced by alternate Lasso models is shown on the top axis. Panel (b) shows increasing R^2 on the lower x-axis (eventually indicating over-fitting to the training data) for alternate regression models where the number of predictors (upper x-axis) is allowed to increase. Both panels indicate the 'sweet spot' of low mean square error, moderately high R^2 value, and parsimony in the number of predictors selected.

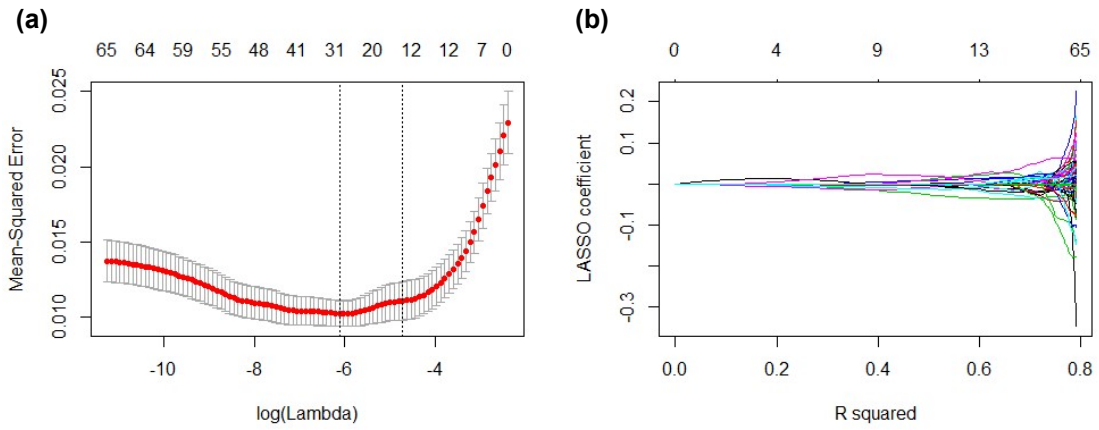


Fig. S9. Alternate Lasso model parameters for the 'Moderate majority' cluster – based on internal cross-validation of *data-split <u>two</u>* training data.

Panel (a) shows mean square error on the y-axis, the log of the penalisation parameter is shown on the lower x-axis, and the increasing number of non-zero predictors introduced by alternate Lasso models is shown on the top axis. Panel (b) shows increasing R^2 on the lower x-axis (eventually indicating over-fitting to the training data) for alternate regression models where the number of predictors (upper x-axis) is allowed to increase. Both panels indicate the 'sweet spot' of low mean square error, moderately high R^2 value, and parsimony in the number of predictors selected.

Table S1. International classification of diseases 10th revision Australian modification (ICD-10-AM) codes for diseases of the respiratory system

This study has used aggregated data for public hospital admissions with principal diagnoses of diseases of the respiratory system (ICD-10-AM codes J00-J99 [14]), listed below, as reported by the Social Health Atlas of Australia [12, 13]. To provide an indicative breakdown of individual disease cases, separations from all hospitals (public + private) are shown in the right-hand columns [15]. Note: at the time of writing, separation data were not available for public hospitals only.

Sub-category	Disease	Separations (% of total)	
		2011-12	2012-13
Acute upper respiratory infections (J00-J06)			
	Acute nasopharyngitis [common cold] (J00)	186 (<0.05%)	164 (<0.05%)
	Acute sinusitis (J01)	1039 (0.3%)	900 (0.2%)
	Acute pharyngitis (J02)	2364 (0.6%)	2145 (0.5%)
	Acute tonsillitis (J03)	12195 (3%)	11617 (2.9%)
	Acute laryngitis and tracheitis (J04)	447 (0.1%)	406 (0.1%)
	Acute obstructive laryngitis [croup] and epiglottitis (J05)	7308 (1.8%)	5799 (1.5%)
	Acute upper respiratory infections of multiple and unspecified sites (J06)	**14463 (3.6%)**	**13883 (3.5%)**
Influenza and pneumonia (J09-J18)			
	Influenza due to certain identified influenza virus (J09)	687 (0.2%)	184 (<0.05%)
	Influenza due to other identified influenza virus (J10)	2400 (0.6%)	4846 (1.2%)
	Influenza, virus not identified (J11)	1178 (0.3%)	1316 (0.3%)
	Viral pneumonia, not elsewhere classified (J12)	3508 (0.9%)	3798 (1%)
	Pneumonia due to Streptococcus pneumoniae (J13)	1775 (0.4%)	2017 (0.5%)
	Pneumonia due to Haemophilus influenzae (J14)	1151 (0.3%)	1237 (0.3%)
	Bacterial pneumonia, not elsewhere classified (J15)	3862 (1%)	3908 (1%)
	Pneumonia due to other infectious organisms, not elsewhere classified (J16)	No data	No data
	Pneumonia, organism unspecified (J18)	**67507 (16.8%)**	**63791 (16%)**
Other acute lower respiratory infections (J20-J22)			
	Acute bronchitis (J20)	2642 (0.7%)	2338 (0.6%)
	Acute bronchiolitis (J21)	**18347 (4.6%)**	**19152 (4.8%)**
	Unspecified acute lower respiratory infection (J22)	**24099 (6%)**	**23631 (5.9%)**
Other diseases of upper respiratory tract (J30-J39)			
	Vasomotor and allergic rhinitis (J30)	681 (0.2%)	645 (0.2%)
	Chronic rhinitis, nasopharyngitis and pharyngitis (J31)	1115 (0.3%)	1297 (0.3%)
	Chronic sinusitis (J32)	10970 (2.7%)	11065 (2.8%)
	Nasal polyp (J33)	3419 (0.9%)	3313 (0.8%)
	Other disorders of nose and nasal sinuses (J34)	**25627 (6.4%)**	**26089 (6.5%)**
	Chronic diseases of tonsils and adenoids (J35)	**40813 (10.2%)**	**41384 (10.4%)**
	Peritonsillar abscess (J36)	3251 (0.8%)	3283 (0.8%)
	Chronic laryngitis and laryngotracheitis (J37)	84 (<0.05%)	78 (<0.05%)
	Diseases of vocal cords and larynx, not elsewhere classified (J38)	4436 (1.1%)	4516 (1.1%)
	Other diseases of upper respiratory tract (J39)	1752 (0.4%)	1677 (0.4%)
Chronic lower respiratory diseases (J40-J47)			
	Bronchitis, not specified as acute or chronic (J40)	2451 (0.6%)	2115 (0.5%)
	Simple and mucopurulent chronic bronchitis (J41)	41 (<0.05%)	57 (<0.05%)
	Unspecified chronic bronchitis (J42)	420 (0.1%)	429 (0.1%)
	Emphysema (J43)	419 (0.1%)	349 (0.1%)
	Other chronic obstructive pulmonary disease (J44)	**61893 (15.4%)**	**62151 (15.6%)**
	Asthma (J45)	**36176 (9%)**	**35257 (8.8%)**
	Status asthmaticus (J46)	2505 (0.6%)	2267 (0.6%)

Bronchiectasis (J47)	5185 (1.3%)	5388 (1.4%)
Lung diseases due to external agents (J60-J70)		
Coalworker's pneumoconiosis (J60)	25 (<0.05%)	27 (<0.05%)
Pneumoconiosis due to asbestos and other mineral fibres (J61)	132 (<0.05%)	124 (<0.05%)
Pneumoconiosis due to dust containing silica (J62)	35 (<0.05%)	25 (<0.05%)
Pneumoconiosis due to other inorganic dusts (J63)	No data	No data
Unspecified pneumoconiosis (J64)	5 (<0.05%)	6 (<0.05%)
Pneumoconiosis associated with tuberculosis (J65)	No data	No data
Airway disease due to specific organic dust (J66)	No data	No data
Hypersensitivity pneumonitis due to organic dust (J67)	150 (<0.05%)	106 (<0.05%)
Respiratory conditions due to inhalation of chemicals, gases, fumes and vapours (J68)	20 (<0.05%)	16 (<0.05%)
Pneumonitis due to solids and liquids (J69)	9857 (2.5%)	9894 (2.5%)
Respiratory conditions due to other external agents (J70)	236 (0.1%)	230 (0.1%)
Other respiratory diseases principally affecting the interstitium (J80-J84)		
Adult respiratory distress syndrome (J80)	156 (<0.05%)	148 (<0.05%)
Pulmonary oedema (J81)	689 (0.2%)	609 (0.2%)
Pulmonary eosinophilia, not elsewhere classified (J82)	277 (0.1%)	271 (0.1%)
Other interstitial pulmonary diseases (J84)	3594 (0.9%)	3638 (0.9%)
Suppurative and necrotic conditions of lower respiratory tract (J85-J86)		
Abscess of lung and mediastinum (J85)	531 (0.1%)	532 (0.1%)
Pyothorax (J86)	722 (0.2%)	733 (0.2%)
Other diseases of pleura (J90-J94)		
Pleural effusion, not elsewhere classified (J90)	6696 (1.7%)	6773 (1.7%)
Pleural plaque (J92)		
Pneumothorax (J93)	3189 (0.8%)	3274 (0.8%)
Other pleural conditions (J94)	326 (0.1%)	345 (0.1%)
Other diseases of the respiratory system (J95-J99)		
Postprocedural respiratory disorders, not elsewhere classified (J95)	No data	No data
Respiratory failure, not elsewhere classified (J96)	3476 (0.9%)	3683 (0.9%)
Other respiratory disorders (J98)	5514 (1.4%)	5958 (1.5%)

(Separations > 3% annual total are highlighted in bold)

Table S2. Socio-geographic variables

The following socio-geographic variables were used as candidate predictor data, sourced from the Social Health Atlas of Australia dataset [13], except Area of local government area and Population density which were derived from spatial mapping of LGAs [16].

Variable	Units	Reference period
Total population	No. persons	2013
Area of local government area	km^2	N/A
Population density	Persons/km^2	2013
Percent Aboriginal persons	%	2013
Percent Australian-born persons	%	2011
Percent born overseas in predominantly English-speaking countries	%	2011
Percent born overseas in a predominantly non-English-speaking country	%	2011
Percent born overseas in a predominantly non-English-speaking country, resident in Australia for 5 years or more	%	2011
Percent born overseas in a predominantly non-English-speaking country, resident in Australia for less than 5 years	%	2011
*Socioeconomic index	No units	2011
Percent smoking during pregnancy	%	2008 to 2010 (NSW, Qld, SA and ACT), 2009 to 2011 (Vic, WA, Tas), 2006 to 2008 (NT)
Percent smokers (estimated population rates, 18 years and over)	ASR /100	2011-13
Percent high alcohol consumption (estimated population rates, 18 years and over)	ASR/100	2011-13
#Percent overweight persons (estimated population rates, 18 years and over)	ASR/100	2011-13
#Percent obese persons (estimated population rates, 18 years and over)	ASR/100	2011-13

N/A = not applicable, ASR = age standardized rate

*Socioeconomic index was originally called Index of relative socioeconomic disadvantage in the Social Health Atlas of Australia. Note that low values correspond to greater disadvantage or deprivation, while high values correspond to reduced disadvantage (or higher socioeconomic status).

#In our data, obese denotes body mass index > 30, while overweight denotes body mass index 25 to <30.

Table S3. Environmental variables

The following environmental variables provided the basis for candidate predictor data (all listed in Table S5). Broad themes of environmental data are: climate (C), ecological (E), geographic (G), land use (LU), pollution (P), soil parameters (S), vegetation or land cover classes (VLC), and vegetation indices from remote sensing (VRS). Units are displayed for numeric variables, whereas the number of classes (in brackets) are displayed for categorical† variables. Further preparatory steps for gridded environmental variables (i.e. 3 km radius focal calculations for numeric average, and conversion of categorical layers to class proportions and Shannon diversity indices), and subsequent selective transformations, are described in the *Methods* (refer to main article). #Focal calculations were not performed for point-based PM10 emissions (air pollution) data, as no exhaustive mapping was available; instead PM10 sites were intersected within each LGA, expressed as total/area, and averaged for the 2 year period. Additional abbreviations are explained in the Table footnote.

Variable	Theme	Units / (classes†)	Reference period	Data source
#Air pollution: Total industry PM10 emissions (mean 2011-12 and 2012-13)/LGA area	P	kg/km^2	2011-2013	[17, 18]
Air temperature annual mean diurnal range	C	°C	1970-2012	[19]
Air temperature annual mean isothermality	C	°C*100	1970-2012	[19]
Annual mean precipitation	C	mm	1970-2012	[19]
Annual mean temperature	C	°C	1970-2012	[19]
Annual precipitation seasonality (coefficient of variation)	C	N/A	1970-2012	[19]
Annual temperature seasonality (standard deviation*100)	C	°C*100	1970-2012	[19]
Distance to coast	G	Decimal degrees	N/A	Calc.
Ecological land units (combinations of bioclimate and landform)†	E	(39)	2012-2013	[20]
Fire frequency mapping for Australia	P	No. years burnt	1997-2010	[21]
Land cover†	VLC	(34)	2000-2008	[1]
Land use†	LU	(15)	1997-2014	[22]
Major vegetation groups†	VLC	(32)	2012	[23]
Maximum temperature of the warmest month	C	°C	1970-2012	[19]
Mean precipitation of the coldest quarter	C	mm	1970-2012	[19]
Mean precipitation of the driest month	C	mm	1970-2012	[19]
Mean precipitation of the driest quarter	C	mm	1970-2012	[19]

Variable	Theme	Units / (classes†)	Reference period	Data source
Mean precipitation of the warmest quarter	C	mm	1970-2012	[19]
Mean precipitation of the wettest month	C	mm	1970-2012	[19]
Mean precipitation of the wettest quarter	C	mm	1970-2012	[19]
Mean temperature annual range	C	°C	1970-2012	[19]
Mean temperature of the coldest quarter	C	°C	1970-2012	[19]
Minimum temperature of the coldest month	C	°C	1970-2012	[19]
Mean temperature of the driest quarter	C	°C	1970-2012	[19]
Mean temperature of the warmest quarter	C	°C	1970-2012	[19]
Mean temperature of the wettest quarter	C	°C	1970-2012	[19]
Prescott Index	C	N/A	1981-2006	[24]
Soil clay content, in the < 2 mm fraction (0-5 cm)	S	%	2014	[3]
Soil effective cation exchange capacity (0-5 cm)	S	meq/100g	2014	[3]
Soil effective cation exchange capacity (0-5 cm) * Soil erodible fraction * Vegetation fractional cover – mean BS; calculated as a geometric mean (intended to represent possible exposures to high clay and organic matter content soils, weighted by erodibility and soil cover)	S	N/A	2000-2012	Calc.
Soil effective cation exchange capacity (0-5 cm) * Soil erodible fraction; calculated as a geometric mean (intended to represent possible exposures to high clay and organic matter content soils, weighted by erodibility)	S	N/A	2014	Calc.
Soil effective cation exchange capacity (0-5 cm) * Vegetation fractional cover – mean BS; calculated as a geometric mean (intended to represent possible exposures to high clay and organic matter content soils, weighted by soil cover)	S	N/A	2000-2012	Calc.
Soil erodible fraction; calculated using the method of Fryrear *et al.* [2]	S	%	2014	Calc.
Soil erodible fraction * Vegetation fractional cover – mean BS; calculated as a geometric mean (intended to represent possible soil exposures due to erodible and bare soils)	S	N/A	2000-2012	Calc.
Soil organic carbon content, in the < 2 mm fraction (0-5 cm)	S	%	2014	[3]
Soil organic carbon stocks in the top 30 cm	S	$t.ha^{-1}$	2010	[25]
Soil pH (in $CaCl_2$) (0-5 cm)	S	pH	2014	[3]
Soil sand content, in the < 2 mm fraction (0-5 cm)	S	%	2014	[3]
Soil silt content, in the < 2 mm fraction (0-5 cm)	S	%	2014	[3]

Variable	Theme	Units / (classes†)	Reference period	Data source
Species richness (average of all biological species occurrences in 9 pane moving window, where each pane is 0.01 degrees latitude/ longitude, or ~1km^2)	E	No. of species	N/A	[26]
Vegetation FPAR maximum	VRS	Fraction	1981-2011	[27]
Vegetation FPAR mean	VRS	Fraction	1981-2011	[27]
Vegetation FPAR median	VRS	Fraction	1981-2011	[27]
Vegetation FPAR minimum	VRS	Fraction	1981-2011	[27]
Vegetation FPAR standard deviation	VRS	Fraction	1981-2011	[27]
Vegetation fractional cover – maximum BS	VRS	%	2000-2012	[28]
Vegetation fractional cover – maximum NPV	VRS	%	2000-2012	[28]
Vegetation fractional cover – maximum PV	VRS	%	2000-2012	[28]
Vegetation fractional cover – mean BS	VRS	%	2000-2012	[28]
Vegetation fractional cover – mean NPV	VRS	%	2000-2012	[28]
Vegetation fractional cover – mean PV	VRS	%	2000-2012	[28]
Vegetation fractional cover – min BS	VRS	%	2000-2012	[28]
Vegetation fractional cover – min NPV	VRS	%	2000-2012	[28]
Vegetation fractional cover – min PV	VRS	%	2000-2012	[28]
Vegetation fractional cover – standard deviation BS	VRS	%	2000-2012	[28]
Vegetation fractional cover – standard deviation NPV	VRS	%	2000-2012	[28]
Vegetation fractional cover – standard deviation PV	VRS	%	2000-2012	[28]

Abbreviations used: N/A = not applicable; Calc. = calculated; FPAR = Fraction of photosynthetically active radiation (i.e. radiation signal due to living green vegetation); PV = photosynthetic (living green) vegetation; NPV = non-photosynthetic vegetation (e.g. non-living straw, stubble, plant remnants), BS = bare soil.

Note: Summary statistical layers (minimum, mean, median, maximum, standard deviation) for time-series vegetation FPAR and fractional cover products were previously calculated by CSIRO Land and Water for collaborative Australia-wide predictive soil and landscape mapping via the Terrestrial Ecosystem Research Network 'Soil and Landscape Grid of Australia' [29].

Table S4. Socio-geographic clustering of Australian local government areas

The total available number of local government areas (LGAs), as indicated, were used in k-means clustering, however a reduced number [in brackets] were used in subsequent modelling due to reduced availability of public health response and candidate predictor data. Cluster means and interquartile ranges (in brackets) are displayed, with scaled cluster centers beneath[#].

Description	n	Cluster characteristic values		
		Socioeconomic index	Percent Aboriginal persons (%)	Population density (persons/km^2)
Cluster A. 'Moderate majority': The majority of Australian LGAs with moderate socioeconomic status and low to moderate population density	451 [364]	974 (945-1003) 0.1164[#]	6.5 (2.0-7.3) -0.1815[#]	95.8 (0.474-22.5) -0.3306[#]
Cluster B. 'Major cities': Highest population density, highest average wealth LGAs concentrated around major capital cities	72 [62]	1041 (1010-1126) 0.7765[#]	0.8 (0.4-1.0) -0.5066[#]	2978 (2087-3531) 2.272[#]
Cluster C. 'Remote disadvantaged': Lowest population density, highest percent Aboriginal persons, and lowest average socioeconomic status LGAs spanning the arid inland to high rainfall northern Australia.	35 [24]	649 (583-696) -3.097[#]	69.6 (54.0-87.4) 3.380[#]	3.1 (0.029-0.856) -0.4142[#]

Notes:

1. Cluster B 'Major cities' contain LGAs associated with capitals Perth, Adelaide, Melbourne, and Sydney. The capital of Brisbane was not included in this cluster due to larger LGA boundaries being defined in the official LGA dataset in that region, which resulted in lower population densities.

2. Summary descriptive statistics for areas of LGAs ultimately used in the modelling are (note these area distributions are positively skewed):

- Cluster A. 'Moderate majority' (n=364): median 2779 km^2; interquartile range 921-5177 km^2; range 9-93211 km^2
- Cluster B. 'Major cities' (n=62): median 27 km^2; interquartile range 14-54 km^2; range 3-103 km^2
- Cluster C. 'Remote disadvantaged' (n=24): median 43047 km^2; interquartile range 1693-159867 km^2; range 70-320706 km^2

Table S5. Candidate predictors selected (and not selected) from Lasso modelling

From the 10-fold Lasso modelling, mean positive association, mean inverse association, low coefficients of questionable reliability (?), and variables not selected by the Lasso (blank) are denoted below. Variables with mean positive coefficients (+, shaded orange) associate with increased hospital admissions, and those with mean negative coefficients (–, shaded light blue) associate with decreased hospital admissions.

Candidate predictors	**Sign of mean association identified from 10-fold Lasso modelling, by LGA cluster**		
	'Moderate majority'	'Major cities'	'Remote disadvantaged'
Social variables			
Total population		–?	
Area of local government area			
Population density			
Percent Aboriginal persons (logit)	+	+	
Percent Australian-born persons			
Percent born overseas in predominantly English-speaking countries (logit)	–		
Percent born overseas in a predominantly non-English-speaking country			
Percent born overseas in a predominantly non-English-speaking country, resident in Australia for 5 years or more			
Percent born overseas in a predominantly non-English-speaking country, resident in Australia for less than 5 years (logit)		–?	
Socioeconomic index	–	–	
Percent smoking during pregnancy	+	+	
Percent smokers (estimated population rates, 18 years and over)		+	
Percent high alcohol consumption (estimated population rates, 18 years and over)			
Percent overweight persons (estimated population rates, 18 years and over) (logit)	–		
Percent obese persons (estimated population rates, 18 years and over)	+	+	
Environmental variables (3 km focal class proportion^, Shannon diversity index†, mean‡)			
Species richness (log10)^	–		
Diversity of land cover†			
Diversity of land use†	–		
Diversity of ecological land units†			
Diversity of major vegetation groups†	–		
Land cover #1: Proportion of extraction sites^			
Land cover #10: Proportion of rainfed sugar^			
Land cover #11: Proportion of wetlands (logit)^			–
Land cover #12: Proportion of forbs open^			
Land cover #13: Proportion of forbs sparse^			
Land cover #14: Proportion of tussock grass closed^			
Land cover #15: Proportion of alpine grass open^			
Land cover #16: Proportion of hummock grass open^			

Land cover #17: Proportion of sedges open^			
Land cover #18: Proportion of tussock grass open (logit)^		+?	
Land cover #19: Proportion of grassland scattered^			
Land cover #2: Proportion of bare areas^			
Land cover #20: Proportion of tussock grass scattered^			
Land cover #21: Proportion of grassland sparse^			
Land cover #22: Proportion of hummock grass sparse (logit)^			+
Land cover #23: Proportion of tussock grass sparse^			
Land cover #24: Proportion of shrubs closed^			
Land cover #25: Proportion of shrubs open^			
Land cover #26: Proportion of chenopod shrubs open^			
Land cover #27: Proportion of shrubs scattered^			
Land cover #28: Proportion of chenopod shrubs scattered^			
Land cover #29: Proportion of shrubs sparse^			
Land cover #3: Proportion of inland waterbodies^			
Land cover #30: Proportion of chenopod shrubs sparse^			
Land cover #31: Proportion of trees closed^			
Land cover #32: Proportion of trees open (logit)^	−		
Land cover #33: Proportion of trees scattered (logit)^			+
Land cover #34: Proportion of trees sparse^			
Land cover #4: Proportion of salt lakes^			
Land cover #5: Proportion of irrigated cropping^			
Land cover #6: Proportion of irrigated pasture^			
Land cover #7: Proportion of irrigated sugar^			
Land cover #8: Proportion of rainfed cropping^			
Land cover #9: Proportion of rainfed pasture (logit)^		+	
Land use #1: Proportion of nature conservation (logit)^	−		
Land use #11: Proportion of irrigated pastures^			
Land use #12: Proportion of irrigated cropping^			
Land use #13: Proportion of irrigated horticulture^			
Land use #14: Proportion of urban intensive use^			
Land use #15: Proportion of intensive plant and animal production^			
Land use #16: Proportion of rural residential farm infrastructure^			
Land use #17: Proportion of mining and waste^			
Land use #18: Proportion of water^			
Land use #4: Proportion of grazing native vegetation^			
Land use #5: Proportion of production forestry^			
Land use #6: Proportion of grazing modified pastures^			
Land use #7: Proportion of plantation forestry^			
Land use #8: Proportion of dryland cropping^			
Land use #9: Proportion of dryland horticulture^			
Climate: Air temperature mean diurnal range‡			
Climate: Air temperature mean isothermality‡			−
Climate: Maximum temperature of the warmest month‡	+		
Climate: Mean annual precipitation‡			

Climate: Mean annual precipitation seasonality‡			
Climate: Mean annual temperature‡			
Climate: Mean annual temperature seasonality‡			
Climate: Mean precipitation of the coldest quarter‡	–?		
Climate: Mean precipitation of the driest month‡			
Climate: Mean precipitation of the driest quarter‡			
Climate: Mean precipitation of the warmest quarter‡			
Climate: Mean precipitation of the wettest month‡			
Climate: Mean precipitation of the wettest quarter‡			
Climate: Mean temperature annual range‡	+		
Climate: Mean temperature of the coldest quarter‡			
Climate: Mean temperature of the driest quarter‡			
Climate: Mean temperature of the warmest quarter‡			
Climate: Mean temperature of the wettest quarter‡		+	+
Climate: Minimum temperature of the coldest month‡			
Climate: Prescott index‡			
Major vegetation groups (MVG) #1: Proportion of rainforests^			
MVG #10: Proportion of other forests and woodlands^			
MVG #11: Proportion of sparse eucalypt woodlands^			
MVG #12: Proportion of tropical eucalypt woodlands with annual grasses > 2 m^			
MVG #13: Proportion of sparse acacia woodlands^			
MVG #14: Proportion of mallee eucalypt woodlands and shrublands^			
MVG #15: Proportion of tall dense thickets^			
MVG #16: Proportion of acacia shrublands^			
MVG # 17: Proportion of other shrublands^			
MVG # 18: Proportion of heathlands^			
MVG # 19: Proportion of grasslands^			
MVG # 2: Proportion of 2 tall eucalypt forests > 30 m^			
MVG # 20: Proportion of arid spinifex grasslands^			
MVG # 21: Proportion of swampy grasses and sedges (logit)^			–
MVG # 22: Proportion of saltbushes and salt marshes^			
MVG # 23: Proportion of mangroves^			
MVG # 24: Proportion of water^			
MVG # 25: Proportion of cleared vegetation^			
MVG # 26: Proportion of unclassified native vegetation^			
MVG # 27: Proportion of naturally bare^			
MVG # 28: Proportion of sea^			
MVG # 29: Proportion of regrowth^			
MVG # 3: Proportion of eucalypt forests 10-30 m (logit)^	–		
MVG # 30: Proportion of unclassified forest^			
MVG # 31: Proportion of other sparse woodlands^			
MVG # 32: Proportion of sparse mallee eucalypt woodlands and shrublands^			
MVG # 4: Proportion of low eucalypt forests < 10 m^			
MVG # 5: Proportion of eucalypt woodlands (logit)^			–
MVG # 6: Proportion of acacia forests and woodlands (logit)^			–

MVG # 7: Proportion of cypress pine forests and woodlands^			
MVG # 8: Proportion of sheoak forests and woodlands^			
MVG # 9: Proportion of paperbark forests and woodlands^			
Soil effective cation exchange capacity (0-5 cm)‡			
Soil clay content, in the < 2 mm fraction (0-5 cm)‡			
Soil sand content, in the < 2 mm fraction (0-5 cm)‡			
Soil silt content, in the < 2 mm fraction (0-5 cm)‡			
Soil organic carbon content, in the < 2 mm fraction (0-5 cm)‡			
Soil pH (CaCl2) (0-5 cm)‡			
Soil organic carbon stocks in the top 30 cm‡			
Soil effective cation exchange capacity (0-5 cm) * Soil erodible fraction; calculated as a geometric mean‡	–?		
Soil effective cation exchange capacity (0-5 cm) * Soil erodible fraction * Vegetation fractional cover – mean BS; calculated as a geometric mean‡			
Soil effective cation exchange capacity (0-5 cm) * Vegetation fractional cover – mean BS; calculated as a geometric mean‡			
Soil erodible fraction * Vegetation fractional cover – mean BS; calculated as a geometric mean‡			
Soil erodible fraction‡			
Distance to coast‡	+		
Fire frequency ‡			
Vegetation Fraction of photosynthetically active radiation (FPAR) maximum (logit)‡		+	
Vegetation FPAR mean‡			
Vegetation FPAR median‡			
Vegetation FPAR minimum‡			
Vegetation FPAR standard deviation‡			
Vegetation fractional cover – maximum BS (logit)‡			+?
Vegetation fractional cover – maximum NPV‡			
Vegetation fractional cover – maximum PV‡			
Vegetation fractional cover – mean BS‡			
Vegetation fractional cover – mean NPV‡			
Vegetation fractional cover – mean PV‡			
Vegetation fractional cover – minimum BS‡			
Vegetation fractional cover – minimum NPV (logit)‡	+		
Vegetation fractional cover – minimum PV (logit)‡	–?		
Vegetation fractional cover – standard deviation BS‡			+
Vegetation fractional cover – standard deviation NPV‡			
Vegetation fractional cover – standard deviation PV‡			
Ecological land unit (ELU) #1: Proportion of artificial or urban area^			
ELU # 10: Proportion of cool semi dry plains^			
ELU # 11: Proportion of cool wet hills^			
ELU # 12: Proportion of cool wet mountains^			
ELU # 13: Proportion of cool wet plains^			
ELU # 14: Proportion of hot dry hills^			
ELU # 15: Proportion of hot dry mountains^			
ELU # 16: Proportion of hot dry plains^			
ELU # 17: Proportion of hot moist hills^			

ELU # 18: Proportion of hot moist mountains^			
ELU # 19: Proportion of hot moist plains^			
ELU # 21: Proportion of hot semi dry mountains^			
ELU # 22: Proportion of hot semi dry plains^			
ELU # 2: Proportion of cold wet hills^			
ELU # 20: Proportion of hot semi dry hills^			
ELU # 23: Proportion of hot wet hills^			
ELU # 24: Proportion of hot wet mountains^			
ELU # 25: Proportion of hot wet plains^			
ELU # 26: Proportion of snow and ice^			
ELU # 27: Proportion of undefined^			
ELU # 28: Proportion of warm dry hills^			
ELU # 29: Proportion of warm dry mountains^			
ELU # 3: Proportion of cold wet mountains^			
ELU # 30: Proportion of warm dry plains^			
ELU # 31: Proportion of warm moist hills^			
ELU # 32: Proportion of warm moist mountains^			
ELU # 33: Proportion of warm moist plains^			
ELU # 34: Proportion of warm semi dry hills^			
ELU # 36: Proportion of warm semi dry plains^			
ELU # 37: Proportion of warm wet hills^			
ELU # 38: Proportion of warm wet mountains^			
ELU # 39: Proportion of warm wet plains (logit)^	+		
ELU # 4: Proportion of cold wet plains^			
ELU # 40: Proportion of water body^			
ELU # 5: Proportion of cool moist hills^			
ELU # 6: Proportion of cool moist mountains^			
ELU # 7: Proportion of cool moist plains^			
ELU # 8: Proportion of cool semi dry hills^			
ELU # 9: Proportion of cool semi dry mountains^			
Air pollution - total industry PM10 emissions (mean 2011-13)			

Abbreviations: MVG = Major vegetation groups, FPAR = Fraction of Photosynthetically Active Radiation, PV = photosynthetic (living) vegetation, NPV = non-photosynthetic (non-living) vegetation, BS = bare soil, ELU = Ecological land unit.

Table S6. Significance testing of regression coefficients for the 'Moderate majority' cluster—based on *data-split one*.

Lasso regression coefficients and selective inference were evaluated using *data-split one*[#]. Then data for top-ranking predictors (ordered by absolute size of standardized regression coefficients) from *data-split two* were input to standard multiple linear regression software, with results shown†.

Predictor	Lasso[#]		Standard†	
	Coef	p-value	Coef	p-value
Proportion of eucalypt forests 10-30m (logit)	-0.0386	0.3363	-0.0270	0.0055 **
Percent obese persons	0.0366	0.2680	0.0098	0.3528
Socioeconomic index	-0.0325	0.2899	-0.0382	0.0008 ***
Distance to coast	0.0256	0.3944	0.0411	0.0076 **
Proportion of warm wet plains (logit)	0.0241	0.0149 *	0.0190	0.0455 *
Proportion of dryland cropping (logit)	-0.0236	0.9195	-0.0046	0.6692
Mean temperature annual range	0.0217	0.5958	0.0293	0.0583 .
Soil cation exchange capacity x erodible fraction – geometric mean	-0.0204	0.5875	-0.0101	0.2607
Diversity of major vegetation groups	-0.0191	0.6542	-0.0324	0.0033 **
Area of local government area (log10)	0.0187	0.3082	-0.0158	0.2515
Percent English-speaking immigrants (logit)	-0.0185	0.1033	0.0133	0.2415
Percent overweight persons (logit)	-0.0096	0.0486 *	-0.0256	0.0124 *
Percent smoking during pregnancy	0.0051	0.6161	0.0233	0.0218 *
Species richness (log10)	-0.0044	0.5686	-0.0376	0.0239 *

Significance codes: 0–0.001: '***', 0.001–0.01: '**', 0.01–0.05: '*', 0.05– 0.1: '.'

Table S7. Significance testing of regression coefficients for the 'Moderate majority' cluster—based on *data-split <u>two</u>*.

Lasso regression coefficients and selective inference were evaluated using *data-split two*[#]. Then data for top-ranking predictors (ordered by absolute size of standardized regression coefficients) from *data-split one* were input to standard multiple linear regression software, with results shown†.

Predictor	Lasso [#]		Standard†	
	Coef	p-value	Coef	p-value
Maximum temperature of the warmest month	0.0468	0.1105	0.0173	0.2579
Socioeconomic index	-0.0438	0.1748	-0.0511	5e-6 ***
Distance to coast	0.0381	0.1142	0.0355	0.0180 *
Diversity of major vegetation groups	-0.0318	0.0869 .	-0.0134	0.2499
Mean temperature annual range	-0.0295	0.0954 .	0.0229	0.2160
Proportion of warm wet hills (logit)	0.0239	0.6361	-0.0364	0.0099 **
Vegetation fractional cover minimum photosynthetic (logit)	-0.0231	0.5251	0.0205	0.2822
Proportion of open trees (logit)	-0.0202	0.5697	-0.0121	0.3738
Percent smoking during pregnancy	0.0201	0.0206 *	0.0171	0.1345
Percent overweight persons (logit)	-0.0197	0.5545	-0.0126	0.2142
Percent Aboriginal persons (logit)	0.0183	0.3977	-0.0030	0.7843
Proportion of warm wet plains (logit)	0.0179	0.2707	0.0454	0.0002 ***
Mean precipitation of the coldest quarter	-0.0139	0.0930 .	-0.0010	0.9295
Air temperature annual mean isothermality	-0.0135	0.2846	-0.0057	0.6077
Diversity of land use	-0.0077	0.1901	-0.0138	0.1933

Significance codes: 0–0.001: '***', 0.001–0.01: '**', 0.01–0.05: '*', 0.05– 0.1: '.'

Supplementary References

1. Geoscience Australia. 2014 Dynamic Land Cover Dataset V1.0, 27 May 2014.
2. Fryrear DW, Krammes CA, Williamson DL, Zobeck TM. 1994 Computing the wind erodible fraction of soils. *Journal of Soil and Water Conservation* **49**, 183-188.
3. Viscarra Rossel et al. 2014 Soil and Landscape Grid National Soil Attribute Maps (3" resolution) - Release 1. v1. (ed. CSIRO) Data collection. http://www.clw.csiro.au/aclep/soilandlandscapegrid/index.html
4. Wilford J, de Caritat P, Bui E. 2015 Modelling the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. *Geoderma* **259–260**, 81-92. (doi:http://dx.doi.org/10.1016/j.geoderma.2015.05.003)
5. R Core Team. 2015 R: A language and environment for statistical computing. (Vienna, Austria, R Foundation for Statistical Computing.
6. Hijmans RJ. 2015 Raster: Geographic Data Analysis and Modeling. R package version 2.4-20.
7. ESRI. 2014 ArcGIS Desktop: Release 10.2. (Redlands, CA, Environmental Systems Research Institute).
8. Friedman JH, Popescu. 2008 Predictive learning via rule ensembles. The Annals of Applied Statistics 2, 916–954. (doi:10.1214/07-AOAS148)
9. Lin L. 1989 A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **45**, 255-268. (doi:10.2307/2532051)
10. Tibshirani R, Tibshirani R, Taylor J, Loftus J, Reid S. 2016 selectiveInference: Tools for post-selection inference. R package version 1.2.0.
11. Taylor J, Tibshirani RJ. 2015 Statistical learning and selective inference. *Proceedings of the National Academy of Sciences* **112**, 7629-7634. (doi:10.1073/pnas.1507583112)
12. PHIDU. 2015 Social Health Atlas of Australia: Data by Local Government Area, March 2015 release. (Adelaide, Public Health Information Development Unit, Torrens University Australia).
13. PHIDU. 2016 Social Health Atlas of Australia: Data by Local Government Area, May 2016 release. (Adelaide, Public Health Information Development Unit, Torrens University Australia).
14. Australian Consortium for Classification Development. 2015 International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification. https://www.accd.net.au/Icd10.aspx
15. AIHW. 2017 AIHW National Hospital Morbidity Database: Separation statistics by principal diagnosis in ICD-10-AM, Australia, 2011−12 to 2012–13. (Canberra, Australian Institute of Health and Welfare). URL: http://www.aihw.gov.au/hospitals-data/principal-diagnosis-data-cubes/ [Accessed 27-08-2017]
16. ABS. 2011 Local Government Area ASGC Ed 2011 Digital Boundaries in ESRI Shapefile Format (Canberra, Australian Bureau of Statistics).
17. NPI. 2012 2011/12 data within Australia - Particulate Matter 10.0 um from All Sources. (National Pollutant Inventory, Australia). http://www.npi.gov.au/npidata/action/load/download-result/criteria/substance/70/destination/AIR/source-type/ALL/substance-name/Particulate%2BMatter%2B10.0%2Bum/subthreshold-data/Yes/year/2012

18. NPI. 2013 2012/2013 data within Australia - Particulate Matter 10.0 um from All Sources. (National Pollutant Inventory, Australia). http://www.npi.gov.au/npidata/action/load/download-result/criteria/substance/70/destination/AIR/source-type/ALL/substance-name/Particulate%2BMatter%2B10.0%2Bum/subthreshold-data/Yes/year/2013
19. Whitley R, Evans B, Pauwels J, Hutchinson M, Xu T, Han W. 2014 Bioclimatic surfaces: eMAST-R-Package 2.0, 0.01 degree, Australian Coverage Obtained from http://dap.nci.org.au, made available by the Ecosystem Modelling and Scaling Infrastructure (eMAST, http://www.emast.org.au) Facility of the Terrestrial Ecosystem Research Network (TERN, http://www.tern.org.au). (Macquarie University, Sydney, Australia).
20. Sayre et al. 2014 A New Map of Global Ecological Land Units — An Ecophysiographic Stratification Approach. (Association of American Geographers, Washington DC).
21. WA Landgate, North Australia Fire Information Service. 2012 1km 1997-2010 Fire frequency mapping for Australia. http://138.80.128.151/firehistory/
22. ABARES. 2014 Catchment scale land use of Australia March 2014. (Australian Bureau of Agricultural and Resource Economics and Sciences, Department of Agriculture and Water Resources). http://www.agriculture.gov.au/abares/aclump/land-use/data-download
23. Department of the Environment and Energy, A. 2012 Major Vegetation Groups - NVIS Version 4.1 (Albers 100m analysis product). http://www.environment.gov.au/land/native-vegetation/national-vegetation-information-system/data-products#mvg41
24. Gallant J, Austin J. 2012 Prescott Index derived from 1" SRTM DEM-S. v2. (ed. CSIRO). Data collection. (doi:10.4225/08/53EB2D0EAE377)
25. Viscarra Rossel R, Webster R, Bui E, Baldock J. 2014 Baseline map of Australian soil organic carbon stocks and their uncertainty. v2. (ed. CSIRO) Data collection (doi:10.4225/08/556BCD6A38737)
26. Atlas of Living Australia. 2016 Species Richness download. Data collection. http://spatial.ala.org.au/
27. Donohue R, McVicar T, Roderick M. 2013 Fraction of Photosynthetically Active Radiation (FPAR) - derived from Advanced Very High Resolution Radiometer (AVHRR), CSIRO Land and Water algorithm, Australia coverage. (ed. CSIRO). (doi:10.4225/08/50FE0CBE0DD06)
28. CSIRO Land and Water. 2012 Fractional cover - MODIS, CSIRO Land and Water algorithm, Australia coverage. Data collection. http://www.auscover.org.au/xwiki/bin/view/Product+pages/Fractional+Cover+MODIS+CLW+to+v2
29. Grundy et al. 2015 Soil and Landscape Grid of Australia. *Soil Research* **53**, 835-844. (doi:http://dx.doi.org/10.1071/SR15191)