Independent validation of national satellite-based land-use regression models for nitrogen dioxide using passive samplers

Luke D. Knibbs[1*], Craig P. Coorey[2], Matthew J. Bechle[3], Christine T. Cowie[4,5,6], Mila Dirgawati[7], Jane S. Heyworth[7], Guy B. Marks[4,5], Julian D. Marshall[3], Lidia Morawska[8], Gavin Pereira[9], Michael G. Hewson[10]

[1] School of Public Health, The University of Queensland, Herston, QLD 4006, Australia

[2] School of Medicine, The University of Queensland, Herston, QLD 4006, Australia

[3] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA

[4] South Western Sydney Clinical School, The University of New South Wales, Liverpool, NSW 2170, Australia

[5] Ingham Institute of Medical Research, Liverpool, NSW 2170, Australia

[6] Woolcock Institute of Medical Research, University of Sydney, Glebe, NSW 2037, Australia

[7] School of Population Health, The University of Western Australia, Crawley, WA 6009, Australia

[8] International Laboratory for Air Quality and Health, Queensland University of Technology, Brisbane, QLD 4001, Australia

[9] School of Public Health, Curtin University, Perth, WA 6000, Australia

[10] School of Geography, Planning and Environmental Management, The University of Queensland, St. Lucia, QLD 4067, Australia

* Corresponding author (e: l.knibbs@uq.edu.au; ph: +61 7 3365 5409; fax: +61 7 3365 5540)

Abstract

Including satellite observations of nitrogen dioxide (NO_2) in land-use regression (LUR) models can improve their predictive ability, but requires rigorous evaluation. We used 123 passive NO_2 samplers sited to capture within-city and near-road variability in two Australian cities (Sydney and Perth) to assess the validity of annual mean NO_2 estimates from existing national satellite-based LUR models (developed with 68 regulatory monitors). The samplers spanned roadside, urban near traffic (≤100 m to a major road), and urban background (>100 m to a major road) locations. We evaluated model performance using R^2 (predicted NO_2 regressed on independent measurements of NO_2), mean-square-error R^2 (MSE-R^2), RMSE, and bias. Our models captured up to 69% of spatial variability in NO_2 at urban near-traffic and urban background locations, and up to 58% of variability at all validation sites, including roadside locations. The absolute agreement of measurements and predictions (measured by MSE-R^2) was similar to their correlation (measured by R^2). Few previous studies have performed independent evaluations of national satellite-based LUR models, and there is little information on the performance of models developed with a small number of NO_2 monitors. We have demonstrated that such models are a valid approach for estimating NO_2 exposures in Australian cities.

Introduction

Land-use regression (LUR) is frequently used for estimating exposure to outdoor air pollution in epidemiological studies. LUR models use features of the built and natural environment, such as road length, impervious surfaces, and tree cover, to capture spatial variability in pollutant concentrations measured at fixed locations. This allows concentrations at unmeasured locations to be estimated.[1] Several recent studies have shown that the predictive ability of LUR models for nitrogen dioxide (NO_2), quantified as R^2, increases by 2 to 15 percentage points when satellite-observed tropospheric NO_2 is included as a predictor variable.[2-7] These models aim to leverage the best attributes of satellite observations (e.g., large spatial coverage) and LUR models (e.g., local-scale predictors) to improve performance and coverage compared with either technique alone.

The spatial coverage offered by satellite data makes it suitable for national or multi-national applications, and satellite-based LUR models have been developed for the USA,[2,7,8] Canada,[6] Australia,[5] Western Europe,[3] and the Netherlands.[4] A single national satellite model can offer a simpler and consistent way to assign exposures to geographically dispersed study subjects compared with separate non-satellite LUR models for each city, which are costly and time-intensive to develop.[9] Some national models can also offer comparable predictive ability and spatial resolution to city-scale models.[2,7]

LUR models can overfit, particularly when the number of measurement sites is small and the number of potential predictors is large.[10-12] Validation is therefore important for assessing how well they perform when applied beyond the data sets used to develop them (e.g., at the home addresses of subjects in an epidemiological study).[12,13] Numerous LUR validation studies have focused on city-scale models (e.g., [11,14,15]). In contrast, there is little information

on validation of satellite-based national NO_2 models,[2,3,7,8] especially in countries with limited ground-based monitoring.[6] Validation of these models is particularly important because they are implemented at a nation-wide scale, which encompasses a wide range of land-use conditions that may differ from the sites used to develop the models.

In this study, we sought to perform an independent validation of Australian national satellite-based LUR models for NO_2. Through this, we wanted to determine if our models were suitable for estimating residential NO_2 exposures in epidemiological studies. We also aimed to add to the limited literature on satellite-based LUR evaluation by exploring the ability of national models developed with a relatively small number of monitoring sites to predict NO_2 concentrations at sites selected to capture within-city and near-road variability.

Experimental Materials and Methods

Models being evaluated

We previously described our satellite-based LUR models for NO_2,[5] which were developed using data from 68 continuous regulatory chemiluminescence monitors throughout Australia (population = 23.5 million; area = 7.7 million km^2; ~0.3 NO_2 monitors/100,000 persons). Two models using different satellite predictors were developed. One model included the tropospheric column abundance of NO_2 molecules observed by the OMI spectrometer aboard the Aura satellite as a predictor (molecules × 10^{15} per cm^2; 'column model'). The other model included the estimated NO_2 concentration at ground-level (ppb; 'surface model'), based on applying a surface-to-column ratio from the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem). Using eight and nine land-use predictor variables, our column and surface models respectively explained 81% (RMSE = 1.4 ppb) and 79% (RMSE

= 1.4 ppb) and of spatial variability in measured annual mean NO_2 in Australia during 2006-11.

Measurements used for validation

In this study, we sought a data set independent of that used in our LUR models' development to rigorously assess their performance. Because we had previously used most available regulatory air monitoring data for development, we contacted all investigators who had performed NO_2 monitoring as part of epidemiological studies between 2006 and 2014. Our initial inclusion criteria were that: (a) NO_2 had been measured anywhere in Australia provided that repeated, precise coordinates were collected (i.e, to 5 decimal places); (b) measurements ran for at least two weeks, and; (c) a validated measurement method with documented quality assurance procedures was used. We received data from five studies, which, to our knowledge, represented all NO_2 monitoring that met the inclusion criteria. Together, these studies included 174 measurement sites across three of Australia's six states.

After preliminary screening we imposed additional, more stringent, inclusion criteria for the studies. Namely, we required three repeated measurements of 14 days' duration each that spanned different seasons. We aimed to ensure that measurements from different studies captured seasonal variation in NO_2, were of comparable duration, and able to be converted to an estimated annual mean using standard methods. These criteria were informed by the well-described European Study of Cohorts for Air Pollution Health Effects (ESCAPE) protocol for LUR development.[16] Based on this, we excluded two studies comprising 43 measurement sites.

The remaining 131 sites were located in Sydney (87 sites; population = 4.9 million) and Perth (44 sites; population = 2 million), the most and fourth-most populous cities in Australia, respectively (Figure 1). All of the sites were located within the metropolitan area of those two cities, and were selected to capture within-city and near-road variability in NO_2. All NO_2 measurements were performed using passive sampling techniques (Ferm-type sampler and Ogawa sampler). Information on sampling dates, measurement methods, and quality assurance is in Table 1.

Conversion to annual mean NO_2

Because each site was measured over two week periods in different seasons but our models' predictions were for annual mean NO_2, we adjusted the measurements to an estimated annual mean. We did this using the ratio of mean NO_2 measured by regulatory monitors during each measurement period compared with its annual mean.[17,18] We calculated the ratio based on three separate regulatory monitors in each study area. We took that approach to improve the precision of the adjusted annual mean estimate (i.e., the overall mean of adjusted concentrations for each measurement period), as measured by its standard error.[17] The selection criteria for the regulatory monitors and the adjustment process are described in the Supporting Information (pages S3-S12).

Site classification

We classified each site as either: (1) roadside (≤15 m to the centre of a major road), or; (2) urban near traffic (not roadside, but ≤100 m to the centre of a major road), or; (3) urban background (not roadside or urban near traffic; >100 m to the centre of a major road). The 15 m distance threshold was selected to capture sites immediately influenced by vehicle emissions, while the 100 m threshold was selected because it represents the approximate half-

life in the decay of NO_2 away from a road.[19-21] Borderline sites on either side of a distance threshold were manually investigated using Google Earth and Street View before assigning them to a category. We assessed the sensitivity of our analyses to a halving and a doubling of the distance thresholds used for classifying roadside sites (7.5 m, 30 m) and urban near traffic sites (50 m, 200 m). Major roads were defined using transport hierarchy codes supplied by the Public Sector Mapping Agencies.[5,22] We also assessed the effect of changing the definition of a major road on our analyses (Supporting Information, pages S22-S26).

There was only one industrial point source of NO_X within 250 m of a site, based on the Australian National Pollutant Inventory.[23] The site was located 120 m from a hospital that emitted a moderate amount of NO_X per year (~5000 kg), but the main source of NO_2 was more likely to be traffic emissions because it was also a roadside site.

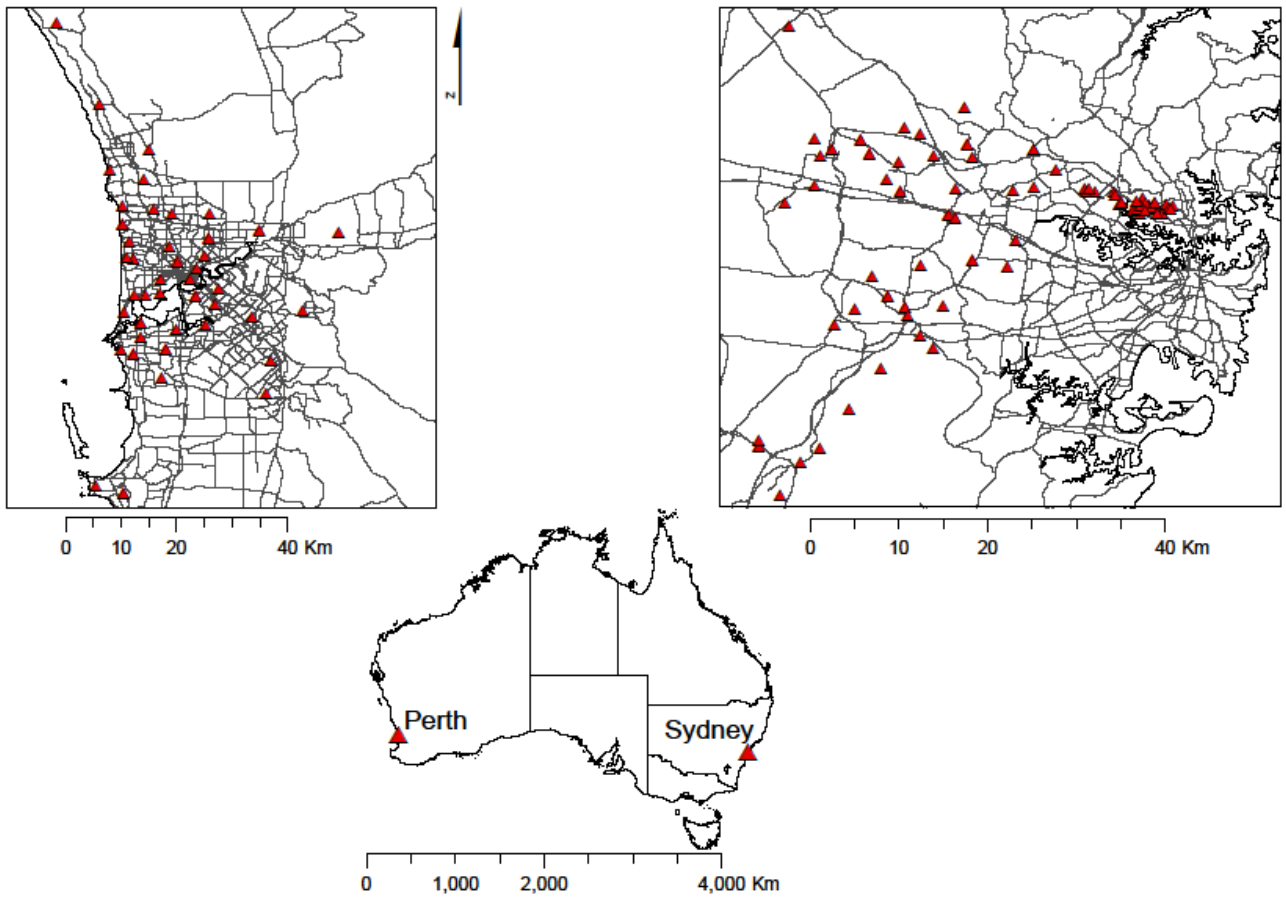


Figure 1. The two Australian cities (Sydney and Perth) where validation measurements were performed. The left panel shows Perth and the right shows Sydney. The map shows the 123 sites used in the main analysis, denoted as red triangles. Major roads are also shown. See Figure S5 in the supporting information for maps of predicted NO_2 in the study areas. The outlines were created using census data published by the Australian Bureau of Statistics and roads were generated from data supplied by the Australian Public Sector Mapping Agencies.[22,37]

170 Table 1. Details of each of the three sampling campaigns used for validation.

	Perth 1	**Sydney 1**	**Sydney 2**
Year	2012	2006-2008	2013-2014
***n* sites**	44	40	47
Site selection	Following ESCAPE protocol[16]	Selected to represent the expected variability of NO_2 in the study area	Following ESCAPE protocol[16]
Sample height	1.5-2 m above ground	2.2 m above ground	2.3-2.4 m above ground
Duration per sample	14 days per sample	14 days per sample	14 days per sample
Timing of samples	1 sample in each of summer, autumn and winter	1 sample in each of summer, winter and spring in each year during 2006-8 (subset of 11-13 sites also sampled in autumn)	1 sample in each of summer (2013) autumn (2014), and winter (2014)
Measurement method	Ogawa sampler[24]	Ferm-type sampler[25]	Ogawa sampler[24]

Analysis	Spectrophotometry based on Saltzman method	Spectrophotometry based on Saltzman method	Spectrophotometry based on Saltzman method
Quality assurance	co-located with chemiluminescence monitors, field blanks, duplicates for each sample	co-located with chemiluminescence monitors, field blanks, duplicates for one in five samples	co-located with chemiluminescence monitors, field blanks, duplicates for one in five samples
Limit of detection	2.0 ppb	0.5 ppb	2.0 ppb
Reference	Dirgawati et al. [26]	Rose et al. [27]	Not yet published

171

172

173

Model predictions

We used our satellite-based LUR models to predict annual mean NO_2 concentrations at each site. Surface and column model predictions were determined for the year in which the validation measurements were performed. Where measurements were done across more than one year, we averaged the predicted NO_2 concentrations to match the measurement period. Measurements from two campaigns (2012 and 2013-14) were performed outside the 2006-11 period used to develop our models. We obtained updated satellite column and surface estimates of NO_2 for those years using our previous methods,[5] and applied them using our existing models. We used all other LUR predictor variables unmodified, based on the assumption that they were unlikely to change substantially over 1-3 years.

We excluded validation sites that had values of one or more LUR predictor variables that were outside the range observed at the 68 regulatory monitoring sites used for model development. We did this to prevent unrealistic predictions, based on the approach of Wang et al.[9,12] Eight sites were excluded, leaving a total of 123 available for validation. We assessed the effect of excluding those sites on our results by comparing them to results with the sites included.

Validation

We used standard methods to validate our LUR models,[12,28] and summarized their performance using an independent validation R^2 (predicted NO_2 regressed on independent measurements of NO_2), the regression slope and 95% confidence intervals, RMSE (absolute and percentage scale), and bias (absolute and fractional). The R^2 we calculated is analogous to a hold-out validation R^2 (HV-R^2),[11-13] except our validation data were a set of unrelated, independent measurements, rather than a subset of model development sites held out for

validation. As such, we refer to our validation metric as R^2 rather than HV-R^2. We performed standard diagnostics on the normality of residuals and their variance. We assessed the spatial correlation of residuals using Moran's *I*.

Because R^2 is based on the correlation between validation measurements and model predictions, it does not reflect their absolute agreement. Therefore, we also calculated a mean-square-error R^2 (MSE-R^2) that took absolute values into account.[10,12,28] MSE-R^2 indicates how well the relationship between measurements and predictions follows a 1:1 line; its derivation is described extensively elsewhere.[10,12,28,29] Using both R^2 and MSE-R^2 can identify LUR model predictions that are well-correlated with measurements but have poor absolute agreement.[10] Unlike R^2, MSE-R^2 can have negative values if the average of measurements leads to a lower MSE than the predictions.[10,12,28,29]

We evaluated LUR model predictions for the entire validation set, by site classification, and by each of the three validation measurement campaigns. We used R version 3.2.2 for all analyses (R Project for Statistical Computing, Vienna, Austria).

Results

NO_2 concentrations

There were 8,177 days of NO_2 measurements performed in total across the 123 validation sites during 2006-2014. Measured NO_2 concentrations adjusted to annual means are summarized in Table 2. Higher concentrations were observed at roadside sites, followed by urban near traffic sites, then urban background sites, and concentrations were higher at sites in Sydney than those in Perth (Table 2). The concentrations we used for validation were slightly higher than those used to develop the LUR models (Table S4). The effects of

changing the definitions used to classify sites on concentration percentiles were minor (Table S5).

Site classification

There were 25 roadside sites, 18 urban near traffic sites, and 80 urban background sites using the standard classification criteria. There was a greater proportion of roadside sites and a smaller proportion of urban background sites used for validation compared with LUR model development, particularly in Perth (Tables S6, S7). However, the percentiles of LUR predictors at validation sites were comparable to the model development sites overall (Table S8). Changing the definitions used to classify sites led to moderate changes in the number of sites in each category (Table S9).

Model validation

Table 3 presents key validation statistics. The surface and column models captured 58% (MSE-R^2 = 51%) and 55% (MSE-R^2 = 52%), respectively, of spatial variability in annual mean NO_2 at the 123 validation sites overall (Figures 2a and 2b). The figures show some evidence of increasing variance of errors with increasing NO_2 concentrations, but plots of predicted NO_2 against residuals did not indicate overt violation of homoscedasticity (Figures S1-S2).

Table 2. Percentiles of annual NO_2 concentrations (ppb) measured at validation sites. * Any negative concentrations following subtraction of field blank values were randomly assigned a value between zero and the limit of detection (2.0 ppb) in the Perth study (see Dirgawati et al. [26]).

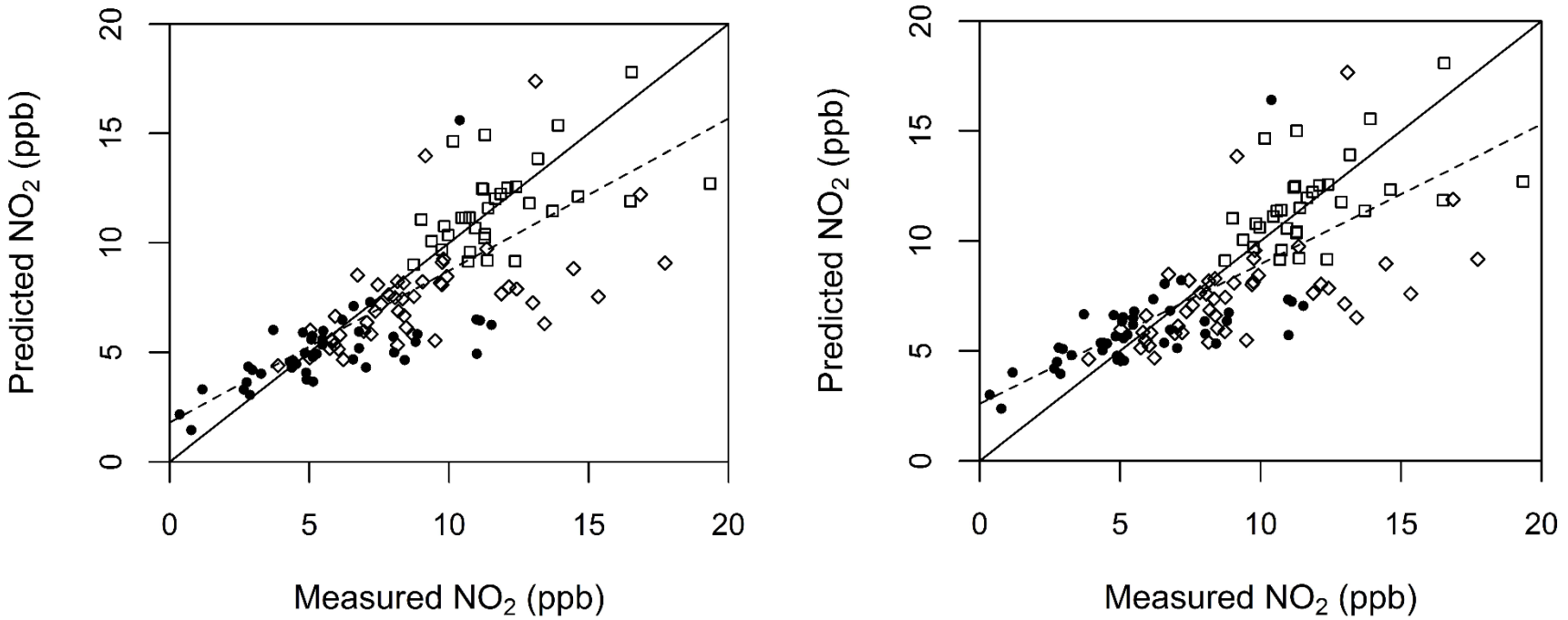
Location	**Min.**	**5th**	**25th**	**50th**	**75th**	**95th**	**Max.**	**Mean**	**S.D.**
overall (*n* = 123)	0.4	2.9	5.9	8.5	11.2	14.6	19.3	8.6	3.7
roadside (*n* = 25)	5.1	5.6	8.0	11.0	13.1	17.6	19.3	11.0	3.9
urban near traffic (*n* = 18)	4.8	4.9	5.8	9.5	11.5	14.8	16.5	9.3	3.6
urban background (*n* = 80)	0.4	2.8	5.2	8.2	10.0	12.4	15.3	7.8	3.3
Sydney (*n* = 80)	3.9	5.8	8.2	9.9	11.9	16.5	19.3	10.2	3.1
Perth (*n* = 43)*	0.4	1.3	4.3	5.1	7.1	11.0	11.5	5.7	2.8

Table 3. Validation statistics for the surface and column models. RMSE = root-mean-square error; FB = fractional bias. Other abbreviations are defined in the main text.

Surface model	**R^2**	**β (95% CI)**	**MSE-R^2**	**RMSE (ppb)**	**RMSE (%)**	**Bias (ppb)**	**FB (-)**
overall (*n* = 123)	0.58	0.69 (0.61, 0.78)	0.51	2.6	29.6	-0.8	-0.10
roadside (*n* = 25)	0.36	0.55 (0.29, 0.81)	-0.18	4.1	37.5	-2.5	-0.26
urban near traffic (*n* = 18)	0.71	0.97 (0.70, 1.24)	0.60	2.2	23.9	-0.2	-0.03
urban background (*n* = 80)	0.68	0.74 (0.65, 0.84)	0.66	1.9	24.6	-0.5	-0.06
urban near traffic + urban background (*n* = 98)	0.69	0.80 (0.71, 0.89)	0.66	2.0	24.5	-0.4	-0.06
Column model							
overall (*n* = 123)	0.55	0.64 (0.55, 0.72)	0.52	2.5	29.5	-0.6	-0.07
roadside (*n* = 25)	0.29	0.47 (0.21, 0.74)	-0.13	4.0	36.7	-2.1	-0.21
urban near traffic (*n* = 18)	0.70	0.91 (0.65, 1.17)	0.64	2.1	22.8	0.1	0.01
urban background (*n* = 80)	0.64	0.67 (0.57, 0.76)	0.64	2.0	25.3	-0.2	-0.03
urban near traffic + urban background (*n* = 98)	0.66	0.73 (0.65, 0.82)	0.65	2.0	24.8	-0.2	-0.02

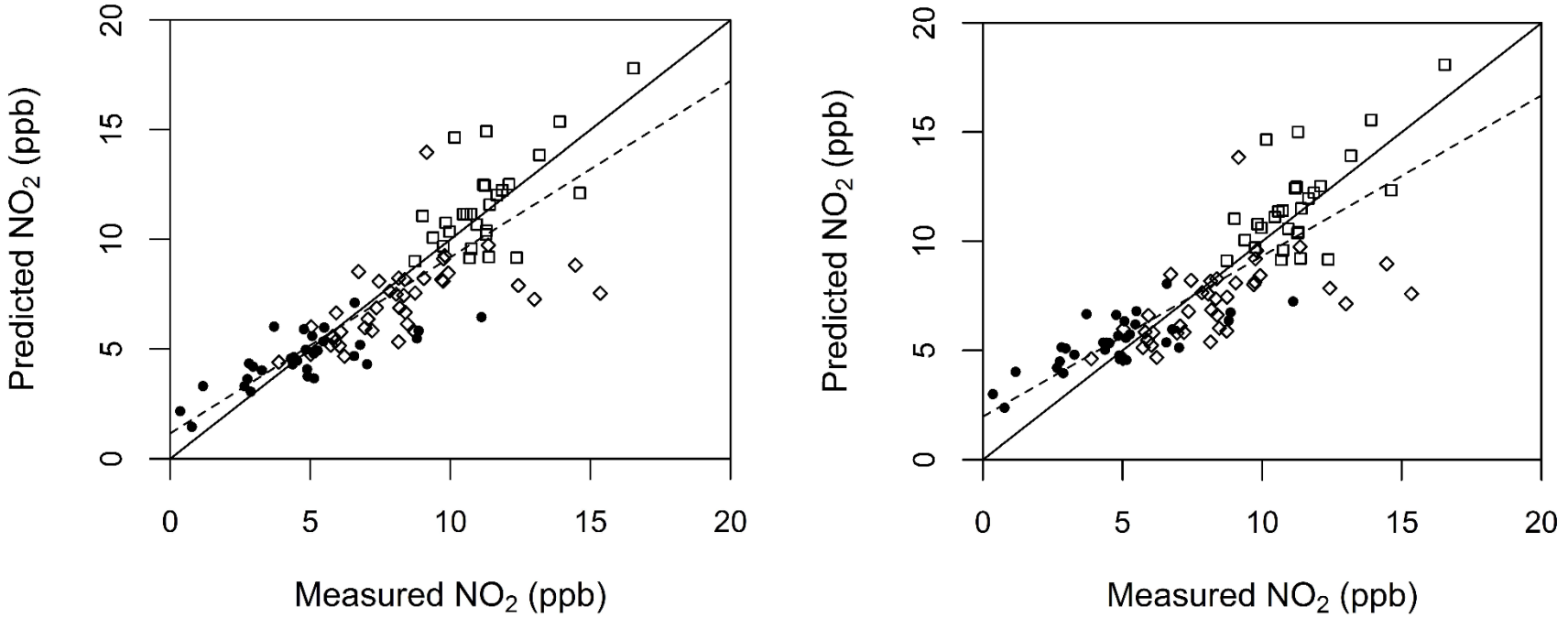
253 The surface model captured 71% (MSE-R^2 = 60%) and 68% (MSE-R^2 = 66%) of spatial
254 variability at urban near traffic and urban background sites, respectively. The column model
255 captured 70% (MSE-R^2 = 64%) and 64% (MSE-R^2 = 64%), respectively. When we combined
256 urban near traffic and urban background sites but excluded the 25 roadside sites, the surface
257 and column models captured 69% (MSE-R^2 = 66%) and 66% (MSE-R^2 = 65%) of spatial
258 variability at the remaining 98 sites, respectively (Figures 3a and 3b). The RMSE and bias of
259 both models was reduced compared with the analysis that included roadside sites. The surface
260 and column models captured 36% (MSE-R^2 = -18%) and 29% (MSE-R^2 = -13%),
261 respectively, of spatial variability at roadside sites.
262
263 *Prediction bias and RMSE*
264 Both models modestly but consistently under-predicted annual mean NO_2, and the column
265 model predicted NO_2 with slightly less bias than the surface model (Table 3). The absolute
266 bias of both models was less than -0.5 ppb for most analyses. Fractional bias was mostly less
267 than -0.10. The absolute RMSE was very similar across both models; approximately 2 ppb
268 (~25% in relative terms). Residuals had an approximately normal distribution and constant
269 variance across all analyses (Figures S1-S4). There was no evidence of spatial correlation
270 among residuals (Table S10).
271
272 *Sensitivity of results*
273 Moving the distance thresholds used to classify roadside and urban near traffic sites led to
274 similar results to the main analysis (Table S9). Likewise, changing the classification of major
275 roads did not substantially alter the results (Table S9). The results of validation stratified by
276 each of the three measurement campaigns are presented in the Supporting Information (Table
277 S11). The predictive ability of both models was lower than that observed when the data were

pooled across all sampling campaigns. Including the eight sites that had predictors outside the range used to develop the models resulted in comparable R^2 values, but lower MSE-R^2 values (Table S12). That finding supported the decision to exclude the sites.



302

303 Figure 2a (left) and 2b (right). Measured *vs*. predicted annual mean NO_2 at 123 validation sites (roadside, urban near traffic, and urban
304 background combined) for the surface (2a) and column (2b) models. The dashed line is the line of best fit (see Table 3 for fit statistics). The solid
305 line is the line of agreement. Symbols denote different measurement campaigns: solid circles = Perth 1; hollow squares = Sydney 1; hollow
306 diamonds = Sydney 2.



307

308 Figure 3a (left) and 3b (right). Measured *vs*. predicted annual mean NO_2 at 98 urban near traffic and urban background validation sites combined
309 for the surface (3a) and column (3b) models. The dashed line is the line of best fit (see Table 3 for fit statistics). The solid line is the line of
310 agreement. Symbols denote different measurement campaigns: solid circles = Perth 1; hollow squares = Sydney 1; hollow diamonds = Sydney 2.

Discussion

Key results and comparison to other studies

Validation of LUR models with data not used in their development is the optimum method for quantifying how well they perform.[12] In this study, we used a large independent set of NO_2 measurements in two Australian cities (n = 123 sites) that was not available at the time of model development to assess the ability of our national satellite-based LUR models (n = 68 sites) to capture within and near-road variability. We previously used five-fold cross-validation with five replications to validate our models due to the scarcity of long-term regulatory NO_2 data in Australia.[5] The model R^2 was 79% (RMSE = 19%) and 81% (RMSE = 19%), respectively, for the surface and column models. Here, we found that our surface and column models explained 69% (RMSE = 25%) and 66% (RMSE = 25%), respectively, of spatial variation in measured annual mean NO_2 at urban near traffic and urban background validation sites combined (n = 98).

Excluding roadside sites, which are discussed in a separate section below, we observed a decrease in R^2 from model development to independent validation of between 10 and 15 percentage points. Bechle et al.[7] assessed their satellite-based LUR for NO_2 in the USA by varying the proportion of sites held out from 10 to 95%. With approximately 70 sites for development and 300 sites for validation, both the model build R^2 (median ~80%) and decrease in R^2 when validated (approximately 10 percentage points) were consistent with what we observed here and in our previous study.[5] Our results also agree with those reported by Wang et al.[12] for a Dutch national, non-satellite LUR for NO_2 developed with 70 sites.

The R^2 decrease we found was less than that described by Hystad et al.[6] for their Canadian national satellite-based LUR for NO_2. They found an average decrease from model

development to independent validation at 618 sites of 34 percentage points (73% *vs.* 39%). Because of the diverse siting of validation sites in their study, its results are more comparable with our overall validation results at 123 sites (i.e., including roadside sites). In that analysis, we observed a decrease in R^2 of 21 and 26 percentage points for the surface and column models, respectively. The smaller reduction in R^2 in this study might reflect the reduced number of sites we used for validation, or the standard criteria we used for repeat measurements and annual adjustment at validation sites to capture seasonal variation in NO_2, which Hystad et al.[6] did for some, but not all, of their sites. It might also reflect that their model had fewer variables (4 predictors *vs.* 8 and 9 predictors in our models) and was not geared towards detecting emissions attributable to heavy industry and biomass combustion, which the authors noted may have affected their results.

Relevance of LUR validation to epidemiological studies

LUR models that have higher out-of-sample R^2 (i.e., between 3 and 16 percentage points lower than model R^2) introduce substantially less attenuation in health effect estimates (from 1% to 14%).[30] The attenuation due to models with lower out-of-sample R^2 (i.e., between 16 and 74 percentage points lower than model R^2) ranges from 9 to 57%, depending on the number of predictors and sites used to develop the model.[30] In the present study, we observed a relatively modest decrease in R^2 from model build through to validation at urban near traffic and urban background sites (10 to 15 percentage points), which was consistent with that in other comparable studies, as outlined above.

Recent work has shown that LUR models with higher independent validation R^2 values produce larger effect estimates than those with lower R^2 values when applied to the association between NO_2 and forced viral capacity (FVC) in children.[13] Model performance

evaluated using leave-one-out-cross-validation (LOOCV) had a much weaker correlation with effect estimates, which underscores the importance of independent validation to determine the utility of LUR models in health studies.[13] Our results demonstrate that the national satellite-based LUR models can be used to estimate with reasonable accuracy the annual mean NO_2 exposures of people living in the metropolitan parts of Australia.

The absolute agreement between pollutant measurements and LUR model predictions is important when models are used to assign exposures in epidemiological studies.[10] Because we aimed to determine if our models were fit for this purpose, we assessed absolute agreement using MSE-R^2. We observed between one and three percentage points difference in R^2 and MSE-R^2 values for urban near traffic and urban background sites combined, and between three and seven percentage points for all sites combined. The differences we found was mostly comparable to those reported by Wang et al.[12] and Basagãna et al.[10] in their European studies. The consistency we observed between R^2 and MSE-R^2 demonstrates that in addition to being correlated, predicted and measured NO_2 also showed similar absolute agreement.

Improving the accuracy of LUR model predictions does not always improve health effect estimates.[29,31] This has been demonstrated when the variability in an LUR predictor is smaller at the measurement sites used to develop the model than the locations to which it will be applied. In turn, this leads to an increase in classical-like measurement error associated with estimating the predictor, which increases bias in the effect estimate compared with a model that has a lower R^2 but less classical error.[29] Such findings illustrate that careful attention needs to be paid to the characteristics of the sites used to develop LUR models versus those they are applied to. In this study, we demonstrated that the percentiles of predictors at

386 validation sites were well-matched to the model development sites (Table S8), and both sets
387 of sites were generally consistent with the ~350,000 census block centroids across Australia
388 (Table S8, [5]). This suggests that our models can be applied to a range of geographic settings
389 within Australia.
390
391 *Surface vs. column model performance*
392 Our surface and column models had similar R^2, MSE-R^2 and RMSE values (Table 3), which
393 agrees with our original model development results.[5] The column model had slightly lower
394 absolute and fractional bias compared with the surface model. We previously reported that
395 column models are a more straightforward and less time-consuming approach, which do not
396 require the simulation of surface-to-column ratios that the surface model does.[5] Since then,
397 Bechle et al.[7] also found that models using tropospheric NO_2 columns performed slightly
398 better than those using surface estimates in a national LUR for the USA. The validation we
399 have described here confirms that column-based NO_2 LUR models for Australia offer a
400 simpler alternative to surface-based models.
401
402 *Performance at roadside sites*
403 The predictive ability of our models at roadside sites (n = 25) was markedly reduced and
404 prediction error increased compared with urban near traffic and urban background sites. The
405 R^2 at roadside sites was 36% (RMSE = 4.1 ppb [38%]) and 29% (RMSE = 4.0 ppb [37%])
406 for the surface and column models, respectively, indicating some correlation between
407 roadside measurements and predictions. The MSE-R^2 values were negative in both cases,
408 indicating poor absolute agreement and that the mean of measurements performed better than
409 model predictions in terms of MSE. Both models under-predicted at roadside locations, with
410 bias of -2.5 ppb and -2.1 ppb for the surface and column models, respectively.

411

412 Our satellite-based LUR models were developed using ambient regulatory monitors, which
413 are deliberately sited away from hotspots like roads. Although the roadside sites used for
414 validation had predictors within the range observed at ambient sites, there was a higher
415 proportion of roadside sites in the validation compared with development data; 20% versus
416 3%, respectively (Table S6). This is a likely explanation for the lower predictive performance
417 at roadside sites. Also, our models were developed for all of Australia and did not include
418 traffic density data because they are not available nationally. We instead used road length
419 data, and the lower predictive ability at roadside sites is probably partially due to the
420 difficulty associated with capturing the variability in NO_2 associated with complex, highly
421 trafficked locations.[32]

422

423 We previously geocoded the residential addresses of 15,000 Australian women randomly
424 selected from Australia's universal healthcare database. We found that the median distance to
425 a major road was 296 m in that cohort, where 84% of women lived in the major cities and
426 inner regional areas of Australia.[33] Moreover, 5.7% of women lived within our definition of a
427 roadside location (≤15 m from a major road), while 8.5% of women lived ≤30 m from a
428 major road. Here, we were mainly interested in the ability of our models to predict at a
429 typical residential address in an epidemiological study, most of which are unlikely to be
430 located immediately proximate to a major road. Our models' performance at roadside
431 locations is therefore less influential on decisions about implementing them in health
432 studies.[12,32]

433

434

435

436 *Limitations*

437 Our study has some important limitations. The validation data we used came from two
438 Australian cities, Sydney and Perth, while the models we sought to validate had national
439 coverage. Sydney and Perth combined (6.9 million people) account for 29% of the Australian
440 population, but it is possible that our validation sites may be less representative of other
441 areas. However, the values of LUR model predictors at our validation sites were largely
442 consistent with those at ~350,000 Australian census block centroids across the country (Table
443 S8), suggesting that the sites are appropriate for validating a national model. Our sites were
444 all located in the metropolitan part of the two cities, which means that validation was not
445 possible in rural and remote parts of Australia. Over 70% of Australians live in major cities,
446 and more than 85% of the population live in urban areas, making Australia one of the world's
447 most urbanized countries.[34] We therefore focused our models' validation on the locations
448 where they will be applied most frequently.

449

450 Although our LUR models were developed using continuous regulatory chemiluminescence
451 monitors we validated them using data from Ferm-type and Ogawa passive samplers.
452 However, these methods have consistently been shown to correlate and agree well for the two
453 week measurement periods we used.[25,35,36]

454

455 Our main analysis only included validation sites that had predictors within the range used to
456 develop our satellite-based LUR models. We did this to prevent unreasonably high or low
457 predictions.[9,12] This means that the predictive performance we observed holds for situations
458 where the predictors are within the models' development range.[10,12] Options for assigning
459 exposures to out-of-range sites in epidemiological studies have been discussed by Wang et
460 al.[12]

In summary, we capitalized on the availability of a large number of NO_2 measurements performed in Australia using standard passive sampling methods, which were not available at the time we built our LUR models. We used almost double the number of sites to validate our models (n = 123) as we used to develop them (n = 68). Our results add to the scant literature on independent validation of national satellite-based LUR models for NO_2, particularly those developed using a relatively small ground-based monitoring network. Our models captured up to 69% of spatial variability in annual mean NO_2 at independent urban near traffic and urban background validation sites, and up to 58% at all validation sites (including roadside sites). Our findings indicate that satellite-based LUR models provide a valid, consistent, and cost-effective method for assigning NO_2 exposures, even when the number of sites available to develop them is limited. Based on the results, we will use the models to estimate residential NO_2 concentrations in a national study of children's respiratory health.

Supporting Information

Tables S1-S12: adjustment to annual mean NO_2; NO_2 concentration percentiles; LUR model development and validation site information; percentiles of predictors at development and validation sites and census block centroids; site classification effects; spatial correlation results; validation results by sampling campaign; effects of excluding sites.

Figures S1-S5: predicted NO_2 vs. residuals; Q-Q plots of residuals; predicted NO_2 in 2008 for Sydney and Perth.

Acknowledgements

L.D.K. acknowledges an NHMRC Early Career (Australian Public Health) Fellowship (APP1036620). G.P. acknowledges a Sidney Sax Fellowship (APP1052236) and project

grants (APP1099655 and APP1047263) from the NHMRC. J.H. acknowledges an NHMRC project grant (APP1003589). C.T.C. acknowledges funding from the Clean Air Research Programme through the Commonwealth Department of Environment, Water, Heritage and the Arts for NO_2 sampling during 2006-2008. Please contact the corresponding author to obtain LUR model predictions for research purposes.

References

1. Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D., A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008,** *42*, (33), 7561-7578.
2. Novotny, E. V.; Bechle, M. J.; Millet, D. B.; Marshall, J. D., National Satellite-Based Land-Use Regression: NO2 in the United States. *Environ. Sci. Technol.* **2011,** *45*, (10), 4407-4414.
3. Vienneau, D.; de Hoogh, K.; Bechle, M. J.; Beelen, R.; van Donkelaar, A.; Martin, R. V.; Millet, D. B.; Hoek, G.; Marshall, J. D., Western European Land Use Regression Incorporating Satellite- and Ground-Based Measurements of NO2 and PM10. *Environ. Sci. Technol.* **2013,** *47*, (23), 13555-13564.
4. Hoek, G.; Eeftens, M.; Beelen, R.; Fischer, P.; Brunekreef, B.; Boersma, K. F.; Veefkind, P., Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmos. Environ.* **2015,** *105*, 173-180.
5. Knibbs, L. D.; Hewson, M. G.; Bechle, M. J.; Marshall, J. D.; Barnett, A. G., A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environ. Res.* **2014,** *135*, 204-211.

6. Hystad, P.; Setton, E.; Cervantes, A.; Poplawski, K.; Deschenes, S.; Brauer, M.; van Donkelaar, A.; Lamsal, L.; Martin, R.; Jerrett, M.; Demers, P., Creating National Air Pollution Models for Population Exposure Assessment in Canada. *Environ. Health Perspect.* **2011,** *119*, (8), 1123-1129.
7. Bechle, M. J.; Millet, D. B.; Marshall, J. D., National Spatiotemporal Exposure Surface for NO2: Monthly Scaling of a Satellite-Derived Land-Use Regression, 2000–2010. *Environ. Sci. Technol.* **2015,** *49*, (20), 12297-12305.
8. Young, M. T.; Bechle, M. J.; Sampson, P. D.; Szpiro, A. A.; Marshall, J. D.; Sheppard, L.; Kaufman, J. D., Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression. *Environ. Sci. Technol.* **2016,** *50*, (7), 3686-3694.
9. Wang, M.; Beelen, R.; Bellander, T.; Birk, M.; Cesaroni, G.; Cirach, M.; Cyrys, J.; de Hoogh, K.; Declercq, C.; Dimakopoulou, K.; Eeftens, M.; Eriksen, K. T.; Forastiere, F.; Galassi, C.; Grivas, G.; Heinrich, J.; Hoffmann, B.; Ineichen, A.; Korek, M.; Lanki, T.; Lindley, S.; Modig, L.; Molter, A.; Nafstad, P.; Nieuwenhuijsen, M. J.; Nystad, W.; Olsson, D.; Raaschou-Nielsen, O.; Ragettli, M.; Ranzi, A.; Stempfelet, M.; Sugiri, D.; Tsai, M. Y.; Udvardy, O.; Varro, M. J.; Vienneau, D.; Weinmayr, G.; Wolf, K.; Yli-Tuomi, T.; Hoek, G.; Brunekreef, B., Performance of multi-city land use regression models for nitrogen dioxide and fine particles. *Environ. Health Perspect.* **2014,** *122*, (8), 843-9.
10. Basagaña, X.; Rivera, M.; Aguilera, I.; Agis, D.; Bouso, L.; Elosua, R.; Foraster, M.; de Nazelle, A.; Nieuwenhuijsen, M.; Vila, J.; Künzli, N., Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmos. Environ.* **2012,** *54*, 634-642.

11. Johnson, M.; Isakov, V.; Touma, J. S.; Mukerjee, S.; Özkaynak, H., Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010,** *44*, (30), 3660-3668.
12. Wang, M.; Beelen, R.; Eeftens, M.; Meliefste, K.; Hoek, G.; Brunekreef, B., Systematic Evaluation of Land Use Regression Models for NO2. *Environ. Sci. Technol.* **2012,** *46*, (8), 4481-4489.
13. Wang, M.; Brunekreef, B.; Gehring, U.; Szpiro, A.; Hoek, G.; Beelen, R., A New Technique for Evaluating Land-use Regression Models and Their Impact on Health Effect Estimates. *Epidemiology* **2016,** *27*, (1), 51-6.
14. Beelen, R.; Voogt, M.; Duyzer, J.; Zandveld, P.; Hoek, G., Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmos. Environ.* **2010,** *44*, (36), 4614-4621.
15. de Nazelle, A.; Aguilera, I.; Nieuwenhuijsen, M.; Beelen, R.; Cirach, M.; Hoek, G.; de Hoogh, K.; Sunyer, J.; Targa, J.; Brunekreef, B.; Künzli, N.; Basagaña, X., Comparison of performance of land use regression models derived for Catalunya, Spain. *Atmos. Environ.* **2013,** *77*, 598-606.
16. ESCAPE Study Manual, ENV.2007.1.2.2.2. European cohort on air pollution, 2008; http://www.escapeproject.eu/manuals/ESCAPE-Study-manual_x007E_final.pdf
17. Hoek, G.; Meliefste, K.; Cyrys, J.; Lewné, M.; Bellander, T.; Brauer, M.; Fischer, P.; Gehring, U.; Heinrich, J.; van Vliet, P.; Brunekreef, B., Spatial variability of fine particle concentrations in three European areas. *Atmos. Environ.* **2002,** *36*, (25), 4077-4088.

18. Lewné, M.; Cyrys, J.; Meliefste, K.; Hoek, G.; Brauer, M.; Fischer, P.; Gehring, U.; Heinrich, J.; Brunekreef, B.; Bellander, T., Spatial variation in nitrogen dioxide in three European areas. *Sci. Total Environ.* **2004,** *332*, (1–3), 217-230.

19. Roorda-Knape, M. C.; Janssen, N. A. H.; de Hartog, J.; Van Vliet, P. H. N.; Harssema, H.; Brunekreef, B., Traffic related air pollution in city districts near motorways. *Sci. Total Environ.* **1999,** *235*, (1–3), 339-341.

20. Gilbert, N. L.; Woodhouse, S.; Stieb, D. M.; Brook, J. R., Ambient nitrogen dioxide and distance from a major highway. *Sci. Total Environ.* **2003,** *312*, (1–3), 43-46.

21. Pleijel, H.; Pihl Karlsson, G.; Binsell Gerdin, E., On the logarithmic relationship between NO2 concentration and the distance from a highroad. *Sci. Total Environ.* **2004,** *332*, (1–3), 261-264.

22. Public Sector Mapping Agencies, Transport and topography product description, v3.6 2013;https://www.psma.com.au/sites/default/files/transport_and_topography_product_description.pdf

23. National Pollutant Inventory (Australia), 2016; http://www.npi.gov.au/

24. Ogawa and Co.; NO, NO_2, NO_x and SO_2 Sampling Protocol Using The Ogawa Sampler, 2006; http://ogawausa.com/wp-content/uploads/2014/04/prono-noxno2so206.pdf

25. Ayers, G. P.; Keywood, M. D.; Gillett, R.; Manins, P. C.; Malfroy, H.; Bardsley, T., Validation of passive diffusion samplers for SO2 and NO2. *Atmos. Environ.* **1998,** *32*, (20), 3587-3592.

26. Dirgawati, M.; Barnes, R.; Wheeler, A. J.; Arnold, A.-L.; McCaul, K. A.; Stuart, A. L.; Blake, D.; Hinwood, A.; Yeap, B. B.; Heyworth, J. S., Development of Land Use

Regression models for predicting exposure to NO2 and NOx in Metropolitan Perth, Western Australia. *Environ. Modell. Softw.* **2015,** *74*, 258-267.

27. Rose, N.; Cowie, C.; Gillett, R.; Marks, G. B., Validation of a Spatiotemporal Land Use Regression Model Incorporating Fixed Site Monitors. *Environ. Sci. Technol.* **2011,** *45*, (1), 294-299.

28. Gulliver, J.; de Hoogh, K.; Hansell, A.; Vienneau, D., Development and Back-Extrapolation of NO2 Land Use Regression Models for Historic Exposure Assessment in Great Britain. *Environ. Sci. Technol.* **2013,** *47*, (14), 7804-7811.

29. Szpiro, A. A.; Paciorek, C. J.; Sheppard, L., Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology* **2011,** *22*, (5), 680-5.

30. Basagaña, X.; Aguilera, I.; Rivera, M.; Agis, D.; Foraster, M.; Marrugat, J.; Elosua, R.; Kunzli, N., Measurement error in epidemiologic studies of air pollution based on land-use regression models. *Am. J. Epidemiol.* **2013,** *178*, (8), 1342-6.

31. Beckerman, B. S.; Jerrett, M.; Serre, M.; Martin, R. V.; Lee, S.-J.; van Donkelaar, A.; Ross, Z.; Su, J.; Burnett, R. T., A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States. *Environ. Sci. Technol.* **2013,** *47*, (13), 7233-7241.

32. Dijkema, M. B.; Gehring, U.; van Strien, R. T.; van der Zee, S. C.; Fischer, P.; Hoek, G.; Brunekreef, B., A Comparison of Different Approaches to Estimate Small-Scale Spatial Variation in Outdoor NO(2) Concentrations. *Environ. Health Perspect.* **2011,** *119*, (5), 670-675.

33. Lazarevic, N.; Dobson, A. J.; Barnett, A. G.; Knibbs, L. D., Long-term ambient air pollution exposure and self-reported morbidity in the Australian Longitudinal Study on Women's Health: a cross-sectional study. *BMJ Open* **2015,** *5*, e008714.

34. Australian Bureau of Statistics, Australian Historical Population Statistics, 2014; http://www.abs.gov.au/ausstats/abs@.nsf/mf/3105.0.65.001

35. Henderson, S. B.; Beckerman, B.; Jerrett, M.; Brauer, M., Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. *Environ. Sci. Technol.* **2007,** *41*, (7), 2422-2428.

36. Eeftens, M.; Beelen, R.; Fischer, P.; Brunekreef, B.; Meliefste, K.; Hoek, G., Stability of measured and modelled spatial contrasts in NO(2) over time. *Occup. Environ. Med.* **2011,** *68*, (10), 765-70.

37. Australian Bureau of Statistics, 2011, Digital Mesh Block Boundaries (Australia), viewed 05 May2016, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage /1270.0.55.001July%202011>