Supplemental for "A Spatial Modeling Approach for Linguistic Object Data: Analysing dialect sound variations across Great Britain"

Shahin Tavakoli,
Statistical Laboratory, University of Cambridge, UK
&
Department of Statistics, University of Warwick, UK,

Davide Pigoli,
Statistical Laboratory, University of Cambridge, UK
&
Department of Mathematics, King's College London, UK,

John A.D. Aston*†
Statistical Laboratory, University of Cambridge, UK,

and

John S. Coleman
Phonetics Laboratory, University of Oxford, UK

May 11, 2018

*The authors gratefully acknowledge support from EPSRC grant EP/K021672/2

†Address for correspondence: Professor John Aston, Statistical laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, CB3 0WB Cambridge, United Kingdom. Email: j.aston@statslab.cam.ac.uk

S1 Square Root of Symmetric Semi-positive Matrices

We give here some useful properties of square root of matrices. The following result states that square root of a symmetric positive semi-definite matrix is unique.

Theorem S1.1 (e.g. Axler (2015))**.** *Let A be a $p \times p$ real matrix. If A is symmetric positive semi-definite, i.e. $A = A^\mathsf{T}$ and $x^\mathsf{T} A x \geq 0, \forall x \in \mathbb{R}^p$, then there exists a unique positive $p \times p$ matrix B such that $A = BB$. The matrix B is called the square root of A, and is denoted by $\sqrt{A}$ or $A^{1/2}$.*

In particular, this tells us that the square root distance between symmetric positive semi-definite matrices is well defined. The following gives a explicit formula for the square root of symmetric positive semi-definite rank one matrices. It's proof follows from direct calculations.

Proposition S1.2. *Let $x \in \mathbb{R}^p, x \neq 0$ and A be a $n \times p$ matrix. Then*

1. $(xx^\mathsf{T})^{1/2} = xx^\mathsf{T}/|x|$
2. $(Axx^\mathsf{T}A^\mathsf{T})^{1/2} = Axx^\mathsf{T}A^\mathsf{T}/|Ax|$ *provided* $Ax \neq 0$.

S2 Technical results and proofs

Lemma S2.1. *The minimizer $\hat{\Omega}(x, \cdot) \in L^2([0,1], \mathcal{S}_p)$ of the following fit criterion:*

$$\sum_{l=1}^{L} K_h\left(\mathrm{d_g}(x, X_l)\right) \int_0^1 d_S^2(\breve{\Omega}_l(t), \hat{\Omega}(x,t))dt, \tag{S2.1}$$

is given by

$$\hat{\Omega}(x,t) = \left[\sum_{l=1}^{L} w_l(x)\sqrt{\breve{\Omega}_l(t)}\right]^2, \tag{S2.2}$$

Proof. Setting $\tilde{w}_l = K_h\left(\mathrm{d_g}(x, X_l)\right)$, using the definition of d_S and permuting the sum and integral in (S2.1), we can rewrite (S2.2) as

$$\int_0^1 \sum_l \tilde{w}_l \left|\left|\left|\sqrt{\breve{\Omega}_l(t)} - \sqrt{\hat{\Omega}(x,t)}\right|\right|\right|^2 dt.$$

This expression is minimized with respect to $\hat{\Omega}$ by minimizing it for each t. Fixing t and writing $y_l = \sqrt{\breve{\Omega}_l(t)}$ and $y = \sqrt{\hat{\Omega}(x,t)}$, and omitting the integral, the fit criterion becomes

$$\sum_l \tilde{w}_l |||y_l - y|||^2.$$

This is just a weighted least-squares problem, whose solution is $y = \sum_l \tilde{w}_l y_l / \left(\sum_l \tilde{w}_l\right)$. Substituting y, y_l back concludes the proof. □

Proof of Proposition 3.1. Without loss of generality, assume that $\mathbb{E}\,X = 0$. By definition, we have

$$\begin{aligned}\operatorname{cov}_{d_S}(X) &= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p} \mathbb{E}\, d_S^2\left(XX^\top, \Omega\right)\\ &= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p} \mathbb{E}\left|\left|\left|\sqrt{XX^\top} - \sqrt{\Omega}\right|\right|\right|^2\\ &= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p} \mathbb{E}\, d_E^2\left(\sqrt{XX^\top}, \sqrt{\Omega}\right).\end{aligned}$$

The minimum is achieved for $\sqrt{\Omega} = \mathbb{E}\sqrt{XX^\top}$, hence $\operatorname{cov}_{d_S}(X) = \left(\mathbb{E}\sqrt{XX^\top}\right)^2 = \mathbb{E}\left[\frac{XX^\top}{|X|}\right]^2$. □

Proof of Proposition 3.2. Without loss of generality, assume that $\mathbb{E}\,X = 0$. By (3.1) of the paper, we have $\operatorname{cov}_{d_S}(AX) = \mathbb{E}\left[\frac{AXX^\top A^\top}{|AX|}\right]^2 = A\ \mathbb{E}\left[\frac{XX^\top}{|AX|}\right]A^\top A\ \mathbb{E}\left[\frac{XX^\top}{|AX|}\right]A^\top$. The proof of the first statement is completed by showing that $\operatorname{cov}_{d_{S,A}}(X)$ must satisfy $A\operatorname{cov}_{d_{S,A}}(X)A^\top = A\ \mathbb{E}\left[\frac{XX^\top}{|AX|}\right]A^\top A\ \mathbb{E}\left[\frac{XX^\top}{|AX|}\right]A^\top$, which follows from an argument similar to the proof of Proposition 3.1. For the special case where $A^\top A = I$, then $|AX| = |X|$ and

$$\operatorname{cov}_{d_S}(AX) = A\ \mathbb{E}\left[\frac{XX^\top}{|X|}\right]^2 A^\top = A\operatorname{cov}_{d_S}(X)A^\top.$$

□

Proof of Proposition 3.3. Let $\tilde{S} = \left(\frac{1}{n}\sum_{i=1}^n \sqrt{(Y_i-\mu)(Y_i-\mu)^\top}\right)^2$, and $S = \operatorname{cov}_{d_S}(Y)$. Recall that $S = \left(\mathbb{E}\sqrt{(Y-\mu)(Y-\mu)^\top}\right)^2$. Let $\phi_x(y) = \sqrt{(x-y)(x-y)^\top}$, for $x, y \in \mathbb{R}^p$. Notice that $\phi_x(y) = (x-y)(x-y)^\top/|x-y|$ if $y \neq x$, and $\phi_x(x) = 0$. Furthermore, it is not difficult to show that ϕ_x is Lipschitz, i.e. $|||\phi_x(y) - \phi_x(y')||| \leq \kappa_p|y-y'|$, where $\kappa_p \geq 0$ does not depend on the value of x, but only on the dimension p. We therefore have

$$\begin{aligned}d_S(\hat{S}, \tilde{S}) &\leq \frac{1}{n}\sum_{i=1}^n \left|\left|\left|\phi_{Y_i}(\overline{Y}) - \phi_{Y_i}(\mu)\right|\right|\right|\\ &\leq \frac{1}{n}\sum_{i=1}^n \kappa_p\left|\overline{Y} - \mu\right|\\ &= \kappa_p\left|\overline{Y} - \mu\right|,\end{aligned}$$

and $d_S(\hat{S}, \tilde{S}) = O_{\mathbb{P}}(n^{-1/2})$.

The proof is completed by showing that $d_S(\tilde{S}, S) = O_{\mathbb{P}}(n^{-1/2})$, which follows from the central limit theorem applied to the random element $\sqrt{(Y-\mu)(Y-\mu)^\top}$. The central limit theorem is indeed applicable here since

$$\begin{aligned}\mathbb{E}\left[\left|\left|\left|\sqrt{(Y-\mu)(Y-\mu)^\top}\right|\right|\right|^2\right] &= \mathbb{E}\left[\left|\left|\left|\frac{(Y-\mu)(Y-\mu)^\top}{|Y-\mu|}\right|\right|\right|^2\right]\\ &= \mathbb{E}\left|Y-\mu\right|^2 < \infty. \square\end{aligned}$$

Proof of Theorem 3.7. By the triangle inequality,

$$d_S(\hat{\Omega}(x,t),\Omega(x,t)) \leq d_S(\hat{\Omega}(x,t),\tilde{\Omega}(x,t)) + d_S(\tilde{\Omega}(x,t),\Omega(x,t)), \tag{S2.3}$$

where $\tilde{\Omega}(x,t)$ is the same as $\hat{\Omega}(x,t)$, but with the sample mean at the observations replaced by the true mean, i.e. $\tilde{\Omega}(x,t) = \left(\sum_{l=1}^{L} w_l(x)\sqrt{\tilde{\Omega}_l(t)}\right)^2$,

$$\sqrt{\tilde{\Omega}_l(t)} = n_l^{-1}\sum_{j=1}^{n_l}\sqrt{(Y_{lj}(t)-m_l(t))(Y_{lj}(t)-m_l(t))^\mathsf{T}} = n_l^{-1}\sum_{j=1}^{n_l}\sqrt{\varepsilon_{lj}(t)\varepsilon_{lj}(t)^\mathsf{T}}$$

and $m_l(\cdot) = \mathbb{E}\, Y_{l1}(\cdot)$.

Let us first look at the first term in (S2.3). Writing $\mathbb{E}_X$ for the expectation conditional on $X_1,\ldots,X_L$, the triangle inequality and Hölder's inequality yield

$$\mathbb{E}_X\, d_S(\hat{\Omega}(x,t),\tilde{\Omega}(x,t)) \leq \sum_{l=1}^{L} w_l(x)\sqrt{\mathbb{E}_X\, d_S^2(\breve{\Omega}_l(t),\tilde{\Omega}_l(t))}.$$

By arguments in the proof of Proposition 3.3 of the paper, we have

$$\mathbb{E}_X\, d_S^2(\breve{\Omega}_l(t),\tilde{\Omega}_l(t)) \leq \kappa_p^2 n_l^{-1}\ \mathbb{E}\,|\varepsilon(x,t)|^2\Big|_{x=X_l} \leq \kappa_p^2 c^{-1} n^{-1} \sup_{x\in\mathcal{E},t\in[0,1]} \mathbb{E}\,|\varepsilon(x,t)|^2,$$

which is non-random, and independent of t. Since $\sum_l w_l(x) = 1$, we get

$$\mathbb{E}_X\, d_S(\hat{\Omega}(x,t),\tilde{\Omega}(x,t)) \leq \frac{\kappa_p}{\sqrt{cn}} \sup_{x\in\mathcal{E},t\in[0,1]} \sqrt{\mathbb{E}\,|\varepsilon(x,t)|^2}$$

Let us now look at the term $\mathbb{E}_X\, d_S(\tilde{\Omega}(x,t),\Omega(x,t)) \leq \sqrt{\mathbb{E}_X\, d_S^2(\tilde{\Omega}(x,t),\Omega(x,t))}$. Since

$$\mathbb{E}_X\, d_S^2(\tilde{\Omega}(x,t),\Omega(x,t)) = \sum_{r,s=1}^{p} \mathbb{E}_X\left(\left[\sqrt{\tilde{\Omega}(x,t)}\right]_{rs} - \left[\sqrt{\Omega(x,t)}\right]_{rs}\right)^2,$$

it is enough to control the mean square error of each coordinate of $\sqrt{\tilde{\Omega}(x,t)}$. Notice that $\sqrt{\tilde{\Omega}(x,t)} = \sum_{l=1}^{L} w_l(x)\sqrt{\tilde{\Omega}_l(t)}$, Therefore we can apply Lemma S2.2 to each coordinate $1 \leq r \leq s \leq p$ (by symmetry), with $Z_l(t) = \left[\sqrt{\tilde{\Omega}_l(t)}\right]_{rs}$. Since $\mathbb{E}_X\sqrt{\tilde{\Omega}_l(t)} = \sqrt{\Omega(X_l,t)}$ and

$$\begin{aligned}
\mathrm{var}_X\left(\left[\sqrt{\tilde{\Omega}_l(t)}\right]_{rs}\right) &= n_l^{-1}\mathrm{var}_X\left(\frac{\varepsilon(X_l,t)_r\varepsilon(X_l,t)_s)}{|\varepsilon(X_l,t)|}\right) \\
&\leq n_l^{-1}\ \mathbb{E}_X\,|\varepsilon(X_l,t)|^2 \\
&\leq \frac{1}{cn}\sup_{x\in\mathcal{E},t\in[0,1]} \mathbb{E}\,|\varepsilon(x,t)|^2
\end{aligned}$$

the Lemma can be applied with $m(x,t) = \sqrt{\Omega(x,t)}$ and $\|\nu\|_\infty \leq (cn)^{-1}\sup_{x\in\mathcal{E},t\in[0,1]} \mathbb{E}\,|\varepsilon(x,t)|^2$. For fixed r,s, the conditional squared bias is bounded by $O_\mathbb{P}(h^2)$ and the conditional variance term is bounded by $O_\mathbb{P}\left(\frac{1}{nLh^2}\right)$, both bounds being uniform in t. The proof is finished by combining these last results. □

Lemma S2.2. *Assume* $(X_l, Z_l(t)) \in \mathcal{E} \times L^2([0,1], \mathbb{R}), l = 1, \ldots, L$ *are i.i.d., with* $X_l \overset{\text{iid}}{\sim} f$, *and assume* $Z_l|X_l$ *are i.i.d. with mean* $\mathbb{E}[Z_l(t)|X_l] = m(X_l, t)$ *and* $\operatorname{var}(Z_l(t)|X_l) = \nu(X_l, t)$, *and that Conditions 3.4, 3.5 of the paper hold. Furthermore, assume*

1. *For each* $t \in [0,1]$, $m(\cdot, t) : \mathcal{E} \to \mathbb{R}$ *is* C^1, *and*

$$\|\nabla_x m\|_\infty := \sup_{x \in \mathcal{E}, t \in [0,1]} \left| \frac{\partial m}{\partial x}(x,t) \right| < \infty,$$

2. f *is a continuous density on* $\mathcal{E}$,

3. $\|\nu\|_\infty := \sup_{x \in \mathcal{E}, t \in [0,1]} \nu(x,t) < \infty$ *for each* $x \in \mathcal{E}$.

Let $\hat{m}(x,t) = \sum_{l=1}^L w_l(x) Z_l(t)$, *where* $w_l(x)$ *is defined in* (3.7) *in the paper. Then for each* x *in the interior of* $\mathcal{E}$, *if* $f(x) > 0$, *we have*

$$|\ \mathbb{E}_X\, \hat{m}(x,t) - m(x,t)| \leq \frac{2\pi \mu_2(K) \|f\|_\infty \|\nabla_x m\|_\infty c_2^2}{c_1^2 f(x)} \left[h + o_{\mathbb{P}}(h) \right], \tag{S2.4}$$

and

$$\operatorname{var}_X(\hat{m}(x,t)) \leq \frac{\|\nu\|_\infty}{Lh^2} \left[\frac{c_2^4}{c_1^4 f^2(x)} + o_{\mathbb{P}}(1) \right] \tag{S2.5}$$

as $L \to \infty, h \to 0$ *such that* $Lh^2 \to \infty$, *where the remainder terms are uniform in* t.

Proof. Without loss of generality, assume that K is renormalized such that $\int_0^\infty K(s) s ds = (2\pi)^{-1}$, and let $\tilde{K}_h : \mathbb{R}^2 \to [0, \infty)$ be defined by $\tilde{K}_h(x) = K(|x|/h)/h^2 = K_h(|x|)$ for $h > 0$. Notice that $\tilde{K}_h$ is a valid density function on $\mathbb{R}^2$ for any $h > 0$, and that it is an approximate identity as $h \to 0$.

We first give a technical result that will be useful, and whose proof follows from standard arguments: for any $\alpha, \beta \geq 0$,

$$\int_{\mathcal{E}} K_h^{1+\alpha}(|x-y|)\, |x-y|^\beta f(y) dy \leq 2\pi \mu_{\beta+1}(K) \|K\|_\infty^\alpha \|f\|_\infty \cdot h^{\beta - 2\alpha}. \tag{S2.6}$$

Recall that $\hat{m}(x,t) = \left[\sum_{l=1}^L K_h(\mathrm{d_g}(x, X_l)) \right]^{-1} \sum_{l=1}^L K_h(\mathrm{d_g}(x, X_l)) Z_l(t)$. First, notice that

$$L^{-1} \left[\sum_{l=1}^L K_h(\mathrm{d_g}(x, X_l)) \right] = \int_{\mathcal{E}} K_h(\mathrm{d_g}(x,y)) f(y) dy + O_{\mathbb{P}} \left(\left[L^{-1} \int_{\mathcal{E}} K_h^2(\mathrm{d_g}(x,y)) f(y) dy \right]^{1/2} \right).$$

By Condition 3.4 of the paper and (S2.6), the stochastic term is of order $O_{\mathbb{P}}(1/\sqrt{Lh^2})$. Concerning the integral, since K_h is an approximate identity as $h \to 0$, approximation theory gives

$$\int_{\mathcal{E}} K_h(\mathrm{d_g}(x,y)) f(y) dy \geq c_2^{-2} \int_{\mathcal{E}} \tilde{K}_{h/c_2}(x-y) f(y) dy = c_2^{-2} f(x) + o(1)$$

as $h \to 0$. Therefore, as $h \to 0, L \to \infty$,

$$\left[L^{-1}\sum_{l=1}^{L} K_h(\mathrm{d_g}(x, X_l))\right]^{-1} \leq \frac{c_2^2}{f(x)} + o_{\mathbb{P}}(1). \tag{S2.7}$$

Let us now look at the bias term. First, notice that $\mathbb{E}_X\, \hat{m}(x,t) = \sum_{l=1}^{L} w_l(x) m(X_l, t)$. Since $x \mapsto m(\cdot, t)$ is C^1, for all $x, y \in \mathcal{E}$, Taylor's theorem yields $m(y,t) = m(x,t) + r(x,y,t)$, where $|r(x,y,t)| \leq \|\nabla_x m\|_\infty |x-y|$. Therefore, using (S2.7),

$$|\ \mathbb{E}_X\, \hat{m}(x,t) - m(x,t)| \leq \left[\frac{c_2^2}{f(x)} + o_{\mathbb{P}}(1)\right] \|\nabla_x m\|_\infty \cdot \left[L^{-1}\sum_{l=1}^{L} K_h(\mathrm{d_g}(x, X_l))|x - X_l|\right]$$

The second term in square brackets is now approximated:

$$\begin{aligned}
L^{-1}\sum_{l=1}^{L} K_h(\mathrm{d_g}(x, X_l))|x - X_l| &\leq c_1^{-2} L^{-1}\sum_{l=1}^{L} K_{h/c_1}(|x - X_l|)|x - X_l| \\
&= c_1^{-2}\int_{\mathcal{E}} \tilde{K}_{h/c_1}(x-y)|x-y| f(y)dy \\
&\quad + c_1^{-2} O_{\mathbb{P}}\left(\left[L^{-1}\int_{\mathcal{E}} K_{h/c_1}^2(|x-y|)|x-y|^2 f(y)dy\right]^{1/2}\right) \\
&\leq c_1^{-2} 2\pi\mu_2(K)\|f\|_\infty \cdot h + O_{\mathbb{P}}(1/\sqrt{L}).
\end{aligned}$$

Combining these results with (S2.7) yields the conditional bias term (S2.4).

Concerning the variance, we have

$$\begin{aligned}
\mathrm{var}_X\left(\hat{m}(x,t)\right) &= \left[L^{-1}\sum_{l=1}^{L} K_h(\mathrm{d_g}(x, X_l))\right]^{-2} \cdot L^{-1}\left[\frac{1}{L}\sum_{l=1}^{L} K_h^2(\mathrm{d_g}(x, X_l))\nu(X_l, t)\right] \\
&\leq \left[\frac{c_2^4}{f^2(x)} + o_{\mathbb{P}}(1)\right] \cdot \|\nu\|_\infty L^{-1} c_1^{-4}\left[\frac{1}{L}\sum_{l=1}^{L} K_{h/c_1}^2(|x - X_l|)\right],
\end{aligned}$$

Where we have used (S2.7). For the term in the second square brackets, we have

$$\begin{aligned}
\left[\frac{1}{L}\sum_{l=1}^{L} K_h^2(|x - X_l|)\right] &= \int_{\mathcal{E}} K_h^2(|x-y|) f(y)dy + O_{\mathbb{P}}\left(\left[\frac{1}{L}\int_{\mathcal{E}} K_h^4(|x-y|) f(y)dy\right]^{1/2}\right) \\
&\leq \|K\|_\infty \|f\|_\infty h^{-2} + O_{\mathbb{P}}(1/\sqrt{Lh^6}),
\end{aligned}$$

where we have used (S2.6). Combining these results yields the conditional variance bound (S2.5). □

The following Lemma gives the approximation error in using the sample total variance in place of the true variance in the estimator of the mean field (3.4) in the paper. Let $\sigma^2(x) = \mathbb{E}\,\|\varepsilon(x)\|^2$.

Lemma S2.3. *Assume*

$$\sup_{x\in\mathcal{E}} \mathbb{E}\,\|\varepsilon(x)\|^4 < \infty, \tag{S2.8}$$

$$c'n \le n_l \le C'n,\ l = 1,\dots,L, \quad \textit{for some constants } c', C',\ \textit{as } n\to\infty, \tag{S2.9}$$

$$\inf_{x\in\mathcal{E}} \sigma^2(x) > 0 \qquad \& \qquad \sup_{x\in\mathcal{E}} \sigma^2(x) < \infty. \tag{S2.10}$$

Let $\hat{m}$ be defined as in (3.4) and let $\check{m}(x) = \sum_{l=1}^{L} \lambda_l(x)\overline{Y}_l$, where

$$\lambda_l(x) = \tilde{\lambda}_l(x)/\sum_{l=1}^{L}\tilde{\lambda}_l(x) \qquad \& \qquad \tilde{\lambda}_l(x) = n_l K_h(\mathrm{d_g}(x, X_l))/\sigma^2(X_l).$$

Then, for fixed L, h,

$$|\hat{m}(x,t) - \check{m}(x,t)| \le O_p(n^{-1/2}) \max_{l=1,\dots,L} |\overline{Y}_l(t)|, \quad \textit{as } n\to\infty.$$

Proof. First, notice that $|\hat{m}(x,t) - \check{m}(x,t)| \le \max_{l=1,\dots,L} |\overline{Y}_l(t)|\,|\sum_l w_l(x) - \lambda_l(x)|$. For the rest of the proof, will drop the x to simplify notation, and write w_l instead of $w_l(x)$. Notice that

$$\left|\sum_l w_l - \lambda_l\right| \le \frac{\sum_l |\tilde{\lambda}_l - \tilde{w}_l|}{s_{\tilde{\lambda}}} + \frac{|s_{\tilde{w}} - s_{\tilde{\lambda}}|}{s_{\tilde{\lambda}}},$$

where $s_{\tilde{w}} = \sum_l \tilde{w}_l$ and $s_{\tilde{\lambda}} = \sum_l \tilde{\lambda}_l$. Using (S2.9), we get that

$$\left|\sum_l w_l - \lambda_l\right| \le (C'/c')^2 \frac{\sum_l |\breve{\lambda}_l - \breve{w}_l|}{s_{\breve{\lambda}}} + (C'/c')^2 \frac{|s_{\breve{w}} - s_{\breve{\lambda}}|}{s_{\breve{\lambda}}}, \tag{S2.11}$$

where the " $\breve{\cdot}$ " entries are the same as the " $\tilde{\cdot}$ " entries, but without the n_ls, i.e. $\breve{w}_l = K_h(\mathrm{d_g}(x, X_l))/\hat{\sigma}^2(X_l)$, and $\breve{\lambda}_l = K_h(\mathrm{d_g}(x, X_l))/\sigma^2(X_l)$. Using (S2.8) and the delta method, we have

$$\breve{\lambda}_l - \breve{w}_l = K_h(\mathrm{d_g}(x, X_l)) \cdot O_{\mathbb{P}}(n^{-1/2}).$$

The first summand in (S2.11) is now bounded:

$$\begin{aligned}\frac{\sum_l |\breve{\lambda}_l - \breve{w}_l|}{s_{\breve{\lambda}}} &\le \frac{O_{\mathbb{P}}(n^{-1/2}) \sum_l K_h(\mathrm{d_g}(x, X_l))}{\sum_l K_h(\mathrm{d_g}(x, X_l))/\sigma^2(X_l)} \\ &= O_{\mathbb{P}}(n^{-1/2}),\end{aligned}$$

where we have used (S2.10). Using the same arguments, we get the same bound on the second summand of (S2.11),

$$\frac{|s_{\breve{w}} - s_{\breve{\lambda}}|}{s_{\breve{\lambda}}} \le O_{\mathbb{P}}(n^{-1/2}).$$

The proof is finished by combining these results. □

S3 Preprocesing of the British National Corpus Data

We describe here in further detail the preprocessing of the sound data extracted from the spoken part of the British National Corpus and analyzed in the paper.

S3.1 Raw Data Preprocessing

First all the segmentation information and all the contextual information were extracted. Then, the list of words for the segmentation and the context were corrected for coding differences (e.g. "they'll" was coded as two separate words "they" and "'ll" in the contextual information files). After this, the segmentation and contextual information were merged together. This was done by matching—within each audio recording file—consecutive groups of words. The algorithm we used looked for a unique sequence of words of length L that perfectly matched between the two sets of words. The algorithm looped through the sequence of utterances (sequence of words pronounce by the same speaker) defined in the contextual XML files, by initially setting L to the minimum of the length of the utterance and 50 (this was chosen for speeding up the matching). If multiple matches were found, L was increased and the search was performed again. If no match was found, L was decreased and the search was performed again. If the algorithm didn't find any match, or if $L > 50$, the algorithm went to the next word in the current utterance (setting $L = 1$). Then L was either increased, respectively decreased, if multiple matches, respectively no match, was found. If $L > 50$, the algorithm was restarted with $L = 1$ but the perfect matching was relaxed to approximate matching using the *optimal string aligment* metric (van der Loo 2014), with distance at most 2.

The result of the preprocessing is a data frame with variables `word`, `begintime`, `endtime`, `textgridfilename`, `index`, `agegroup`, `role`, `sex`, `soc`, `dialecttag`, `age`, `persname`, `occupation`, `dialect`, `id`, `placename`, `activity`, `locale`, `wavfile`, `placenamecleaned` and about 5 million observations (i.e. words). Discriminative information about the speaker is missing for about 2.9% of the words, and information about the location of the recording is missing for about 8.4% of the words.

S3.2 Cleaning

Since the data we analyzed are sounds from noisy recording, we first cleaned the sounds corresponding to the set of words

`class, glass, grass, past, last, brass, blast, ask, cast, fast,pass.` (S3.1)

The following sounds were removed:

1. Sounds with duration outside the interval $[0.2, 1]$ seconds.
2. 400 sounds with the lowest maximal amplitudes.
3. Sounds corresponding to young speakers (selected by taking speakers less than 10 year old and whose median pitch was above a fixed threshold)

To further remove low quality sounds from our analysis, we ranked the sounds $s_1, \ldots, s_N$, for each word w in (S3.1), according to following score,

$$\text{score}_i = \frac{1}{L_i} \sum_{l=1}^{L_i} \left(\check{s}_i(t_l) - \mathbf{1}_{[a(w),b(w)]}(t_l/t_{L_i}) \right)^2 \exp\left(-\mathbf{1}_{[a(w),b(w)]}(t_l/t_{L_i}) \right), \tag{S3.2}$$

where $\check{s}_i(t_l) = \tilde{s}_i(t_l)/\max_{l=1,\ldots,L_i} \tilde{s}_i(t_l)$, $\tilde{s}_i$ is the root mean square amplitude (RMSA) of s_i on a running window of 10 milliseconds, and $a(w), b(w) \in [0,1]$ were chosen by looking at the plot of $\tilde{s}_i$ for a sound of good quality, and correspond roughly to the location of the vowel in the sound. Large values of score_i correspond to noisier sounds. The effect of the exponential factor in (S3.2) is to give higher score to sounds having large RMSA outside the vowel interval, while still penalizing for low RMSA inside the vowel interval. For each word w of our list of words, we then discarded the sounds with the largest 5% scores.

S3.3 Vowel Segmentation and MFCC Extraction

We extracted the MFCCs of all the sounds corresponding to the words in (S3.1), using the software `ahocoder` (`http://aholab.ehu.es/ahocoder/index.html`) with parameter `--CCORD=30 --LFRAME=16`.

In order to extract the MFCC corresponding to the vowel segment of the recording of the words in (S3.1), we performed the following steps. For each word in (S3.1):

1. align the MFCCs of the sounds of the word with respect to the first MFCC coefficient,
2. find the segment of the warped sounds which corresponds to the vowel,
3. extract the corresponding portion on the unwarped MFCCs,
4. recompute all the unwarped MFCCs on a common grid,

S3.4 MFCC alignment

Let us describe more precisely the alignment step in the preprocessing procedure. Let $\text{MFCC}_i(t,m), i = 1, \ldots, N$ denote the MFCCs of the sounds corresponding to the current word w. Recall that $m = 1, \ldots, M$, and assume that the time domains have been linearly rescaled, i.e. $t \in [0,1]$. We first align the curves $\text{MFCC}_i(\cdot, 1), i = 1, \ldots, N$ using the Fisher-Rao metric. This yields warping functions $\gamma_i : [0,1] \to [0,1]$ such that $\text{MFCC}_i(\gamma_i(\cdot), 1), i = 1, \ldots, N$ are aligned. Then we align all the MFCC coefficients of the sound i using the warping γ_i, that is, we set $\widetilde{\text{MFCC}}_i(t,m) = \text{MFCC}_i(\gamma_i(t), m), t \in [0,1], m = 1, \ldots, M$ for all i. The idea is that, after alignment, the temporal location of the vowel would be the same accross all registered MFCCs of a same word, which would make the vowel segmentation much easier.

Once the MFCCs corresponding to a common word w have been aligned, the interval $[a(w), b(w)] \subset [0,1]$ corresponding to the vowel sound was found by manual auditory discrimination. The inverse of the warping functions were then used to compute the interval

$I_i = [\gamma_i^{-1}(a(w)), \gamma_i^{-1}(b(w))]$, which is the vowel interval of the i-th unaligned MFCCs. The interval I_i was then linearly rescaled to $[0, 1]$, yielding the vowel MFCCs

$$\mathrm{MFCC}_i^{\mathrm{vowel}}(t, m) = \mathrm{MFCC}_i\Big((1-t)\gamma_i^{-1}(a(w)) + t\gamma_i^{-1}(b(w)), m\Big), \quad t \in [0, 1] \tag{S3.3}$$

S4 Modeling the Vowel Sound Duration

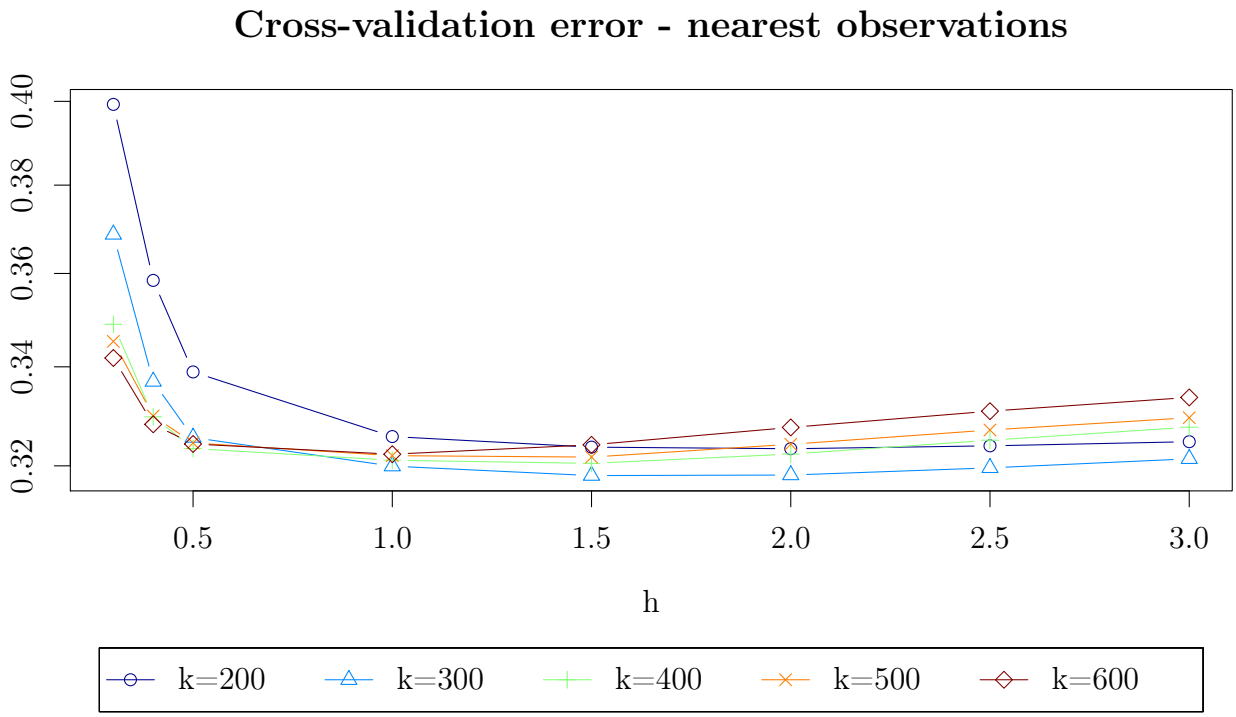
The sound duration of the vowel in the words of the "class" dataset are believed to carry part of the information of the spatial variation of the dialect sounds. However, since the duration cannot capture time dynamics in relative volume, and differences in the vowel quality, the information carried by the vowel duration is a very crude approximation of the vowel sound. This is why the focus of the paper is on the MFCCs of the vowel sounds. We have nevertheless produced a spatial map of the relative duration of the vowel sound (relative to the duration of the word), where the spatial map is obtained by spatial smoothing of the relative durations at each observation location, obtained using a linear mixed model with `observation location`, `word` and `sex` as fixed effects, and `speaker` as random effect. The resulting map is given in Figure S5, together with the projection of the mean MFCC field onto the second principal component. The same spatial smoothing parameters have been used for both maps ($h = 0.5, k = 14$ nearest locations). It can be seen that the two maps are quite correlated (the absolute correlation is 0.66; note that the principal component is defined up to a sign), and therefore the duration information is more or less similar to that obtained by the projection of the MFCC mean field onto the second principal component.

S5 Simulation Study

In order to quantify whether the spatial mean function and the spatial d_S-covariance contain valuable spatial informations, we compare the results obtained in the main paper with a simulation scenario in which all the spatial locations have the same mean and d_S-covariance. We simulate observations from a model with constant mean and constant d_S-covariance,

$$Y_{lj}^* = \mu + \varepsilon_{lj}^*, \quad l = 1, \ldots, L; j = 1, \ldots, n_l \tag{S5.1}$$

where $\mu = \left(\sum_{l,j} Y_{lj}\right) / \sum_l n_l$, ε_{lj}^* were drawn with replacement from $\{\check{\varepsilon}_{lj} : l = 1, \ldots, L; j = 1, \ldots, n_l\}$, $\check{\varepsilon}_{lj} = \hat{\varepsilon}_{lj} - \left(\sum_{l,j} \hat{\varepsilon}_{lj}\right) / \sum_l n_l$, $\hat{\varepsilon}_{lj} = Y_{lj} - \hat{m}(X_l)$, and where $\hat{m}$ is the estimated of the mean MFCC field obtained from the data with tuning parameters $h = 0.5, k = 14$ nearest locations, and n_l is the number of observations at location X_l.



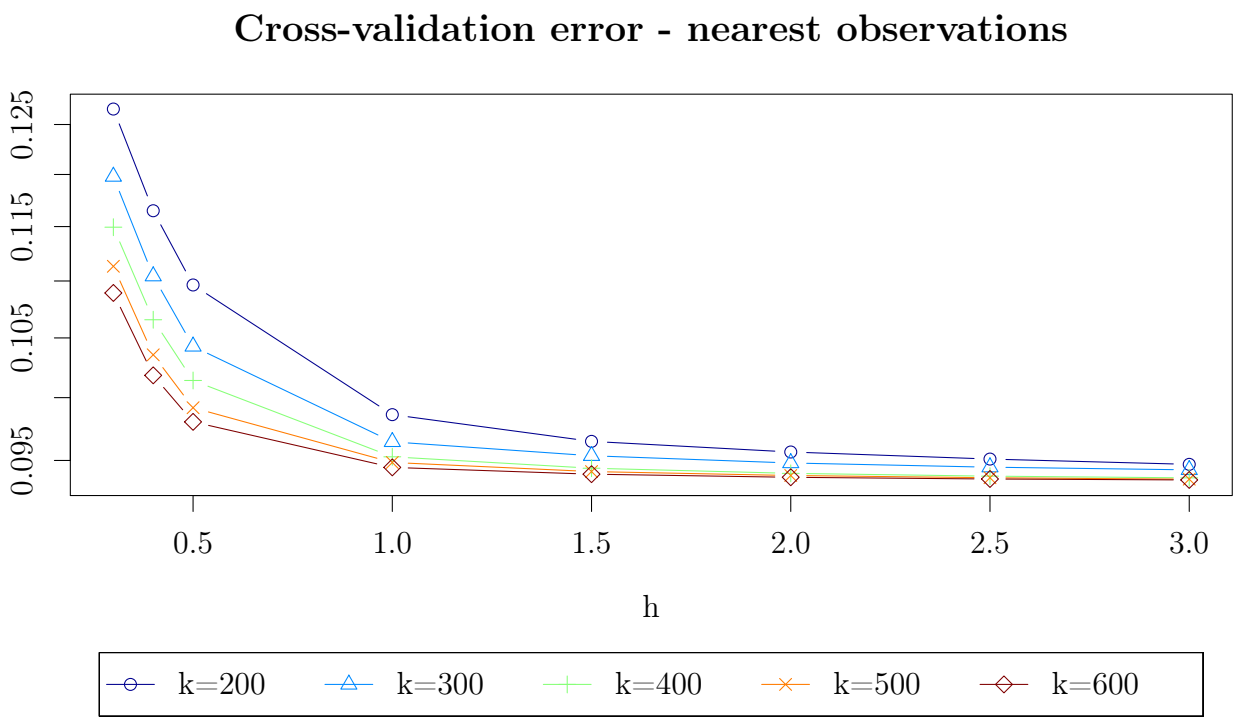


Figure S1: Cross-validation curves of the "class" dataset for the mean MFCC field (top) and the d_S-covariance field (bottom) when the bandwidth is adjusted using the k-th nearest observations.

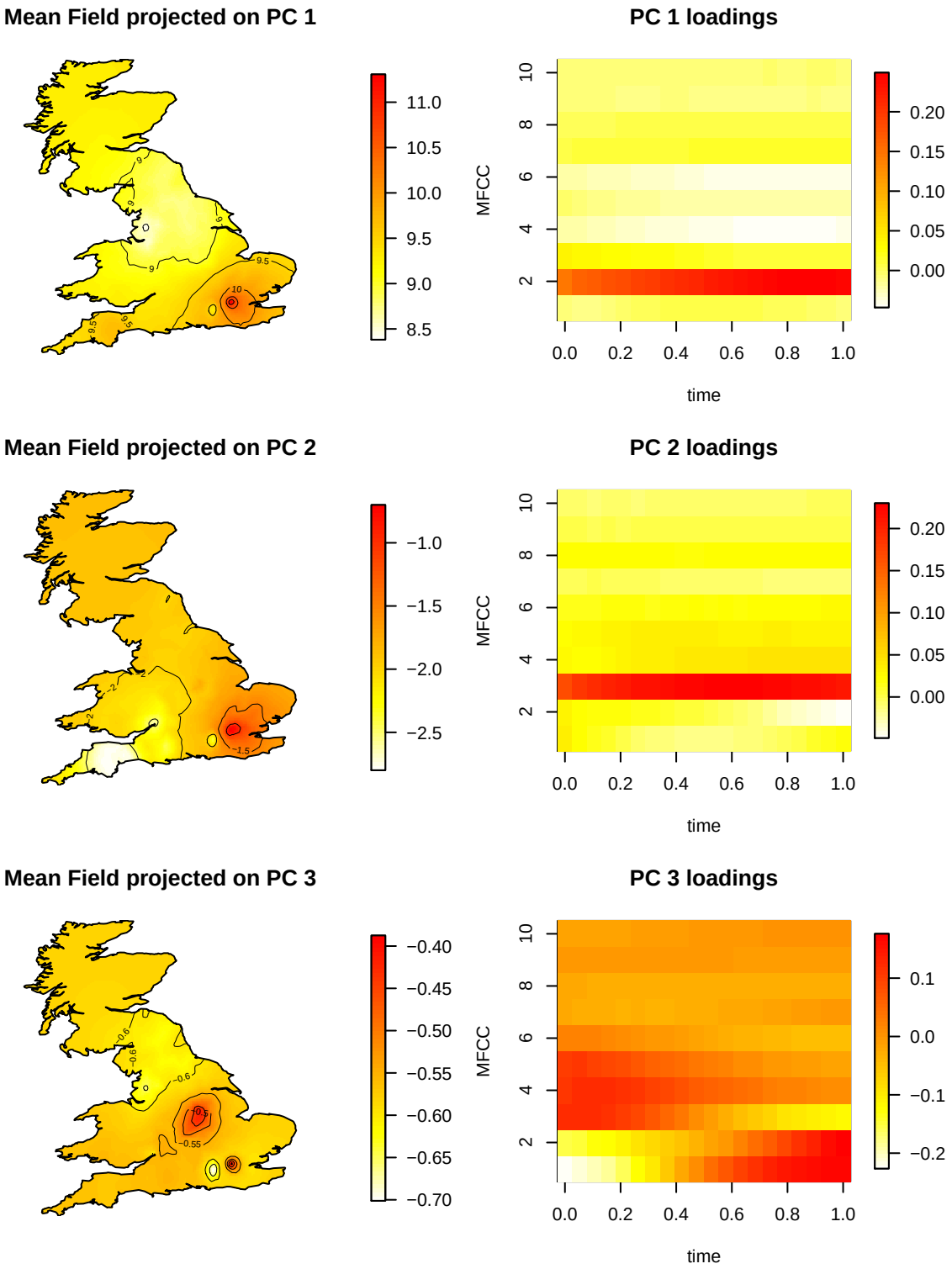


Figure S2: Left: Color maps with contours of the mean smooth MFCC field obtained for the "class" vowel with $h = 1.5$ and $k = 300$th nearest observations (denoted *NO map* in the text), projected onto the first three principal components directions (from top to bottom) of the original data $\{Y_{lj}(t) : l = 1, \ldots, L; j = 1, \ldots, n_l\}$. Right: Colour image representing the projection directions (loadings).

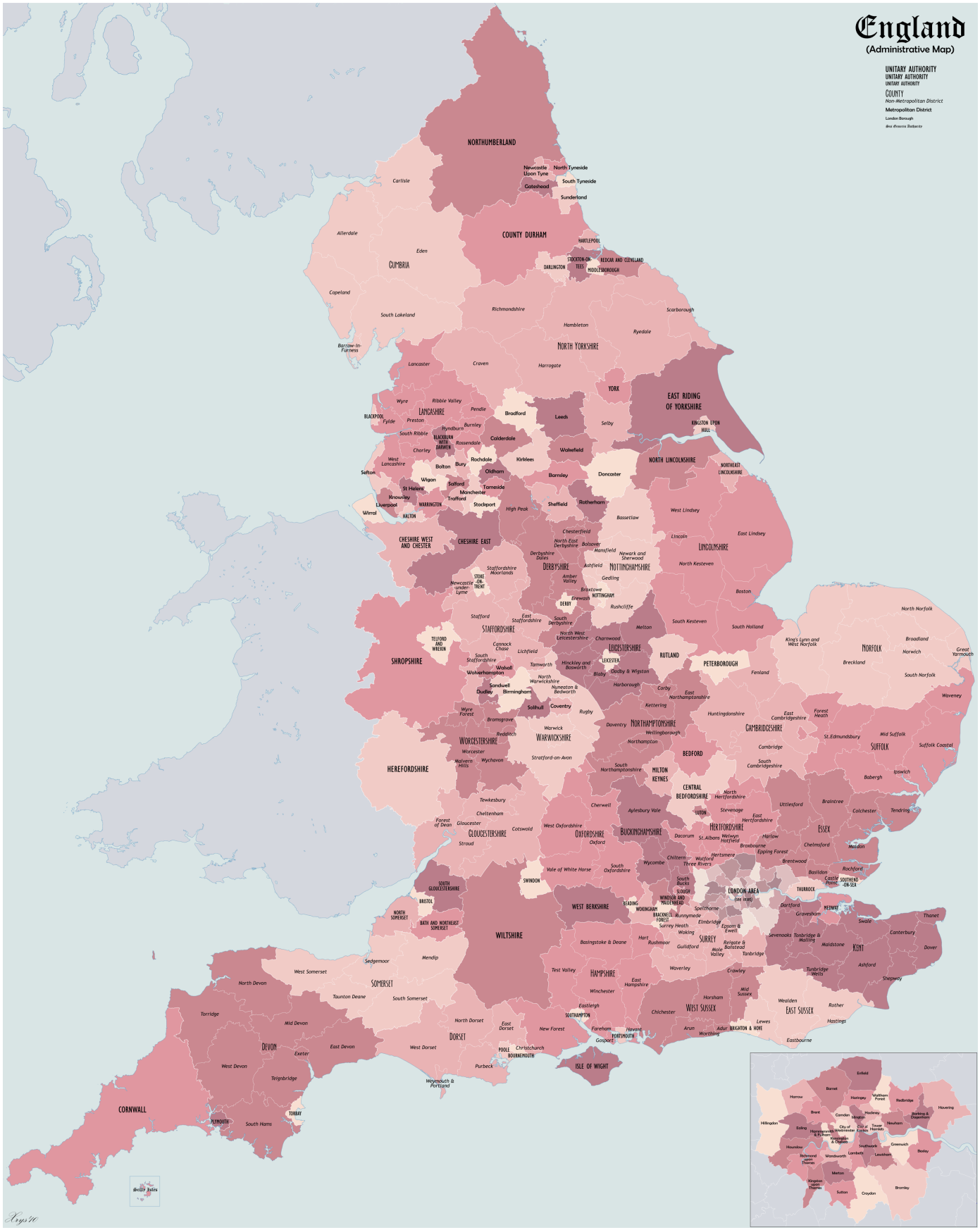


Figure S3: Counties of England. Licenced under the Creative Commons Attribution 3.0 Unported license. Attribution: XrysD.
https://en.wikipedia.org/wiki/File:England_Administrative_2010.png.

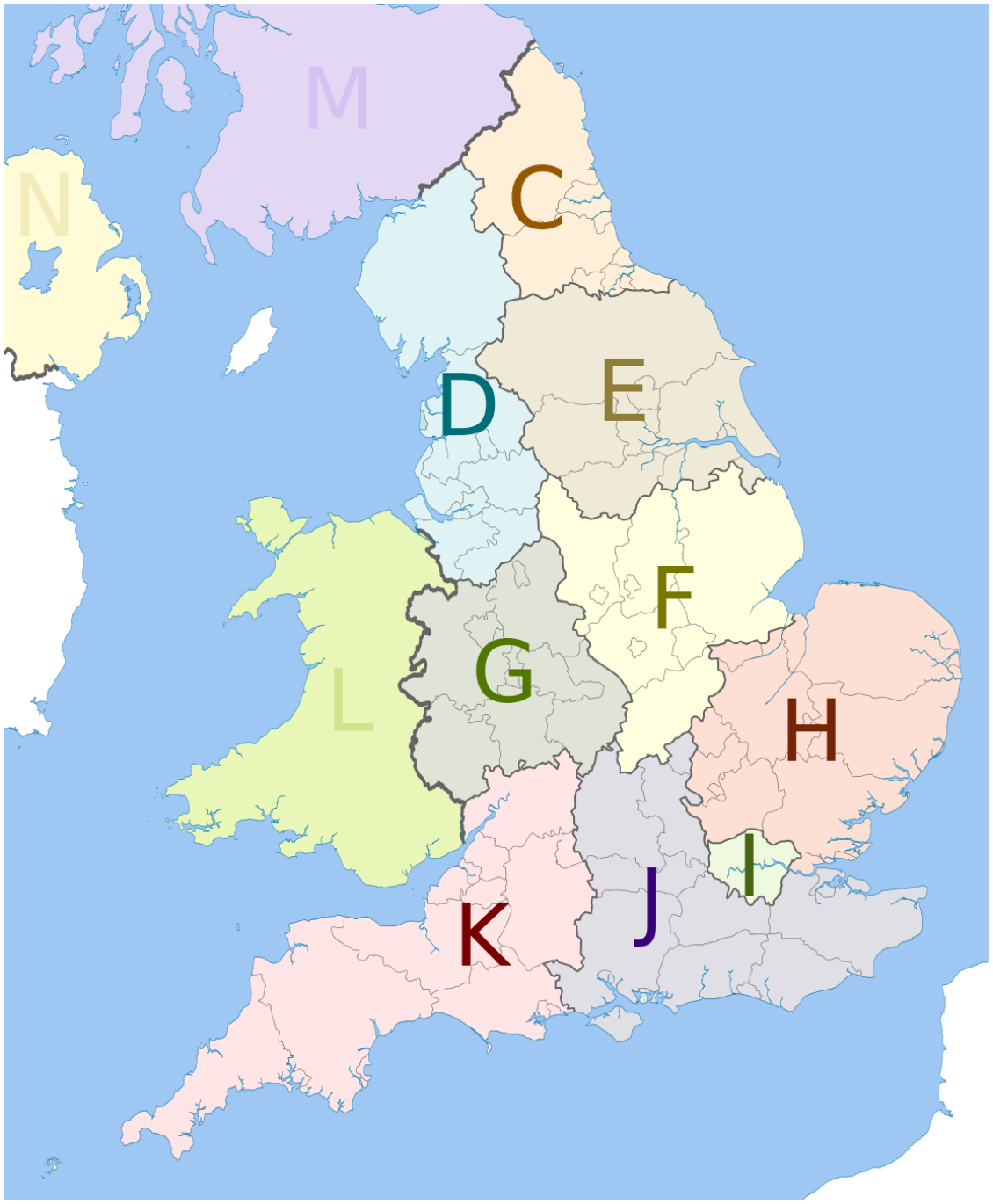


Figure S4: Regions of Great Britain. C = North East England, D = North West England, E = Yorkshire and the Humber, F = East Midlands, G = West Midlands, H = East of England, I = Greater London, J = South East England, K = South West England, L = Wales, M = Scotland. Licenced under the Creative Commons Attribution-Share Alike 3.0 Unported license. Attribution: Dr Greg and Nilfanion.
`https://commons.wikimedia.org/wiki/File:NUTS_1_statistical_regions_of_England_map.svg`.

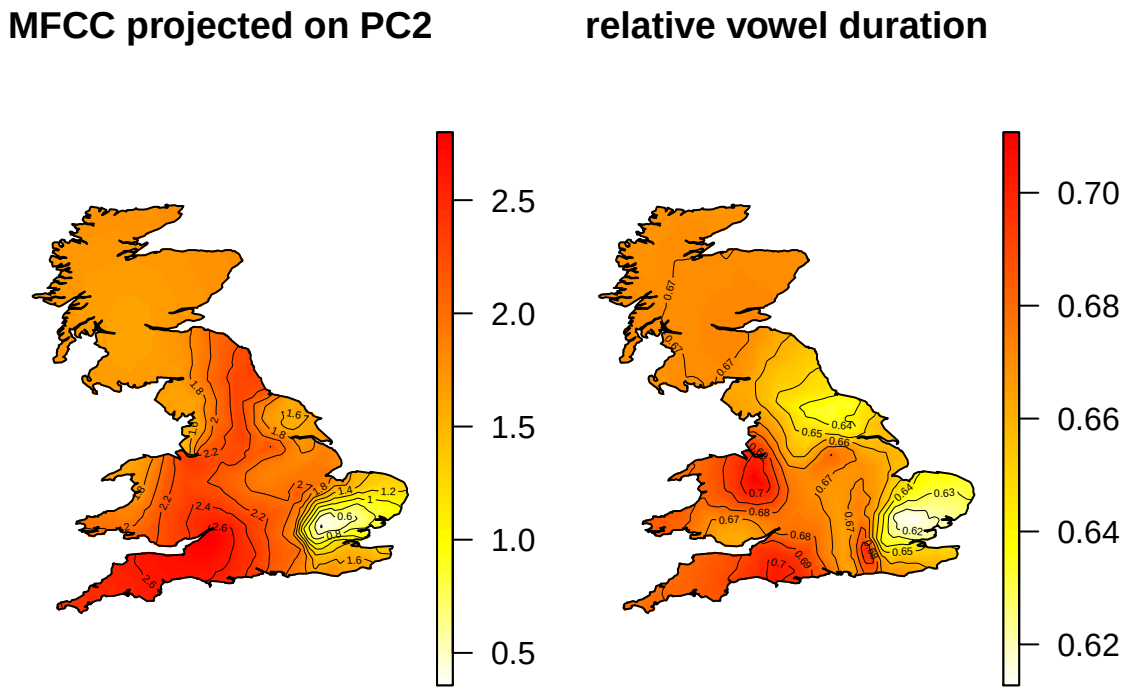


Figure S5: Mean MFCC field projected on PC2 (left) and duration field of the vowel sounds (right). The absolute correlation between the two fields is 0.66.

Notice that although the simulated data is generated under a constant mean model, their estimated d_S-covariance field will be the same as what would be obtained by a model with varying mean, i.e. replacing μ by $\hat{m}(X_l)$ in (S5.1). Indeed, the d_S-covariance field is based on the spatial smoothing of the sample d_S-covariance at each location, defined by

$$\breve{\Omega}_l^*(t) = \left[\frac{1}{n_l}\sum_{j=1}^{n_l}\sqrt{(Y_{lj}^*(t)-\overline{Y}_l^*(t))(Y_{lj}^*(t)-\overline{Y}_l^*(t))^\intercal}\right]^2, \tag{S5.2}$$

where $\overline{Y}_l^* = \sum_j Y_{lj}^*/n_l$. Changing μ in (S5.1) to $\hat{m}(X_l)$ would not change (S5.2), since

$$\begin{aligned} Y_{lj}^* - \overline{Y}_l^* &= (\mu + \varepsilon_{lj}^*) - \sum_i (\mu + \varepsilon_{li}^*)/n_l \\ &= \varepsilon_{lj}^* - \sum_i \varepsilon_{li}^*/n_l \\ &= \left(\hat{m}(X_l) + \varepsilon_{lj}^*\right) - \sum_i (\hat{m}(X_l) + \varepsilon_{li}^*)/n_l. \end{aligned}$$

The d_S-covariance field estimated in each simulations run is therefore the same, regardless of the choice of the mean at each location.

The projections onto PC1-3 are given in Figure S9. If there was no spatial information in the mean field of the BNC dataset, the mean field (projected onto PC1) of the simulated

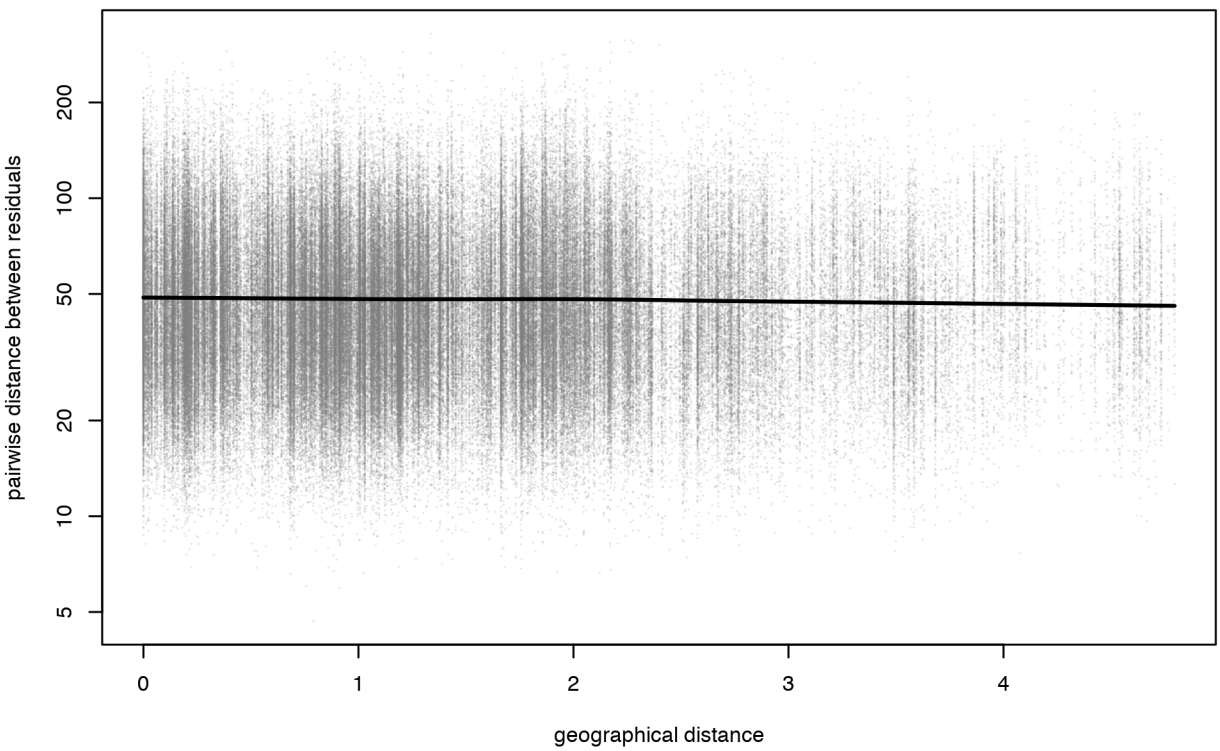


Figure S6: Scatterplot of pairwise distance between residuals (y axis) against their geographical distance (x axis). The black thick line is a robust local linear regression obtained via the R function `lowess`.

data would have the same range of variation as the mean field of the BNC dataset (projected onto PC1). However, the MFCC field of the estimated MFCC field of the simulation has consistently a much smaller range than the smooth field obtained from the BNC dataset over the 100 simulation replicates (the range for the projection on PC1 is [9.2, 9.7] for a realization from (S5.1), as opposed to [7.9, 10.8] for the real data application). This provides evidence in support of spatial structure for the mean field.

S6 An illustration of the advantage of the d_S-covariance

As a motivation for the use of d_S-covariances, here is a one-dimensional example which illustrates the advantages of using them when smoothing spatially under the metric d_S. Suppose you have data $Y_{11}, \ldots, Y_{1m} \overset{\text{iid}}{\sim} \varepsilon(x_1)$ and $Y_{21}, \ldots, Y_{2m} \overset{\text{iid}}{\sim} \varepsilon(x_2)$, where $x_1, x_2 \in \mathbb{R}$ are two points that are equally close to $x_0 \in \mathbb{R}$, and we wish to estimate the co-variation of $\varepsilon(x_0)$. Assume that $\varepsilon(x) \sim N(0, \sigma^2)$ for all $x \in \mathbb{R}$, and that the mean of $\varepsilon(x)$ is known to be equal to zero. If we wish to estimate the parameter $\sigma^2 = \text{var}\,(\varepsilon(x_0))$, then a natural

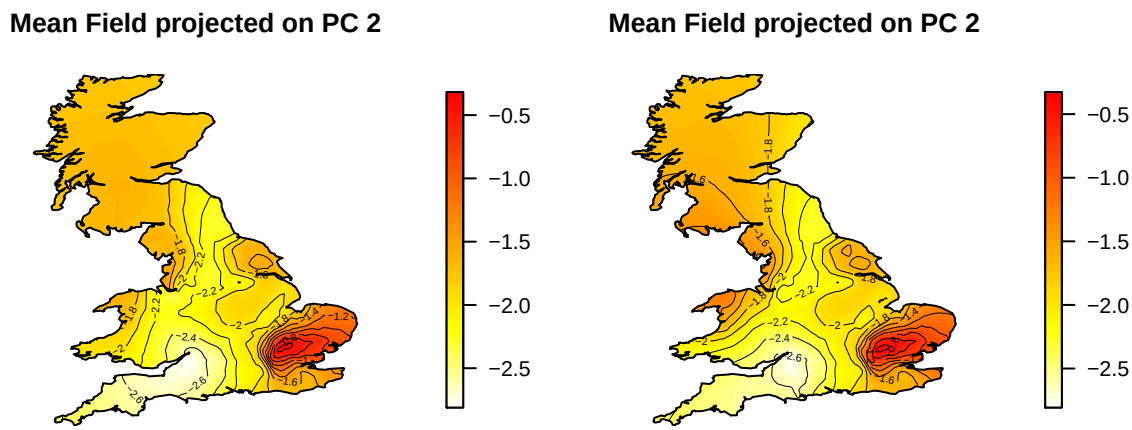


Figure S7: Mean field of the BNC dataset projected on PC2 with computed with geodesic distance (left) and with Euclidean distance (right). Notice the artifacts near the boundaries (the level curves go across the port of Edinburgh when using the Euclidean metric).

estimator is the Fréchet mean of $\hat{\sigma}_i^2 = m^{-1}\sum_{j=1}^m Y_{ij}^2, i = 1, 2$, under d_S, i.e.

$$\hat{\sigma}_*^2 = \left[\left(\sqrt{\hat{\sigma}_1^2} + \sqrt{\hat{\sigma}_2^2}\right)/2\right]^2.$$

But

$$\begin{aligned}
\mathbb{E}\,\hat{\sigma}_*^2 &= \mathbb{E}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/4 + \mathbb{E}\sqrt{\hat{\sigma}_1^2\hat{\sigma}_2^2}/2 \\
&< \sigma^2/2 + \sqrt{\mathbb{E}\,\hat{\sigma}_1^2\hat{\sigma}_2^2}/2 \\
&= \sigma^2/2 + \sqrt{\mathbb{E}(\hat{\sigma}_1^2)\ \mathbb{E}(\hat{\sigma}_2^2)}/2 \\
&= \sigma^2,
\end{aligned}$$

where we have used Jensen's inequality in the second line (which is in this case a strict inequality, since $\hat{\sigma}_1^2\hat{\sigma}_2^2$ is not almost surely constant), and the independence of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in the third line. In other words, $\hat{\sigma}_*^2$ is a biased estimator of σ^2. Furthermore, since $\sqrt{\hat{\sigma}_1^2}$ and $\sqrt{\hat{\sigma}_2^2}$ are both Chi distributed with m degrees of freedom, $\mathbb{E}\sqrt{\hat{\sigma}_*^2} = \sigma\sqrt{2}\Gamma\left((m+1)/2\right)/\Gamma(m/2)$, where Γ is the Gamma function, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$, that is, $\sqrt{\hat{\sigma}_*^2}$ is a biased estimator of $\sqrt{\sigma^2}$. In other words, if one smooths the sample variances using the square-root Euclidean metric, the resulting estimator is biased, even in the square-root space. However, if one wishes to estimate the parameter $\tau = \mathrm{cov}_{d_S}(\varepsilon(x_0)) = [\,\mathbb{E}\,|\varepsilon(x_0)|]^2$, then the natural estimator is the Fréchet mean of

$$\hat{\tau}_i = \left(m^{-1}\sum_{j=1}^m |Y_{ij}|\right)^2, \quad i = 1, 2$$

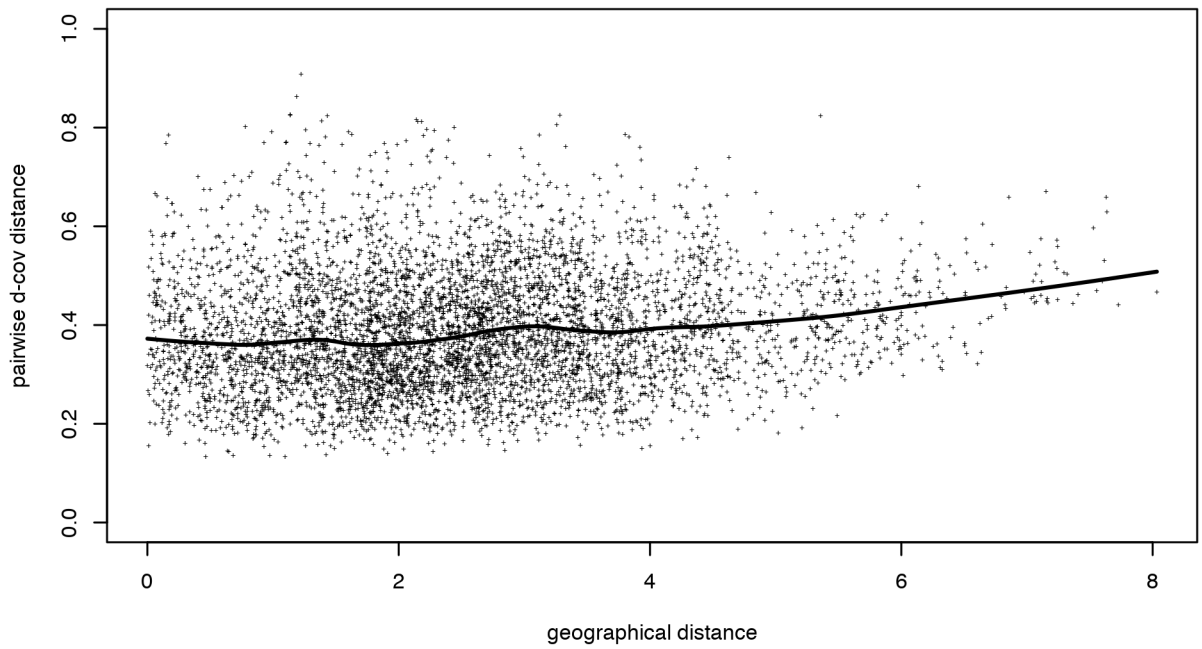


Figure S8: Scatterplot of the distances between the raw d_S-covariances $\breve{\Omega}_l, \breve{\Omega}_k$, and the corresponding geographical distance between X_l, X_k. Notice that the thick line, which represents a robust local linear regression obtained via the R function `lowess`, has a nugget, and is slightly increasing with the geographical distance.

under d_S, that is

$$\hat{\tau}_* = \left[(2m)^{-1} \sum_{j=1}^{m} \left(|Y_{1j}| + |Y_{2j}| \right) \right]^2,$$

which is unbiased in the square-root space, i.e. $\mathbb{E}\sqrt{\hat{\tau}_*} = \sqrt{\tau}$. In conclusion, using the same metric for the spatial smoothing and the definition of the co-variation yields estimators that are less biased than those obtained by using distinct metrics.

S6.1 Comparison of the d-covariance field under the square-root metric and the Euclidean metric

One might raise the question of whether the d_S-covariance field yields results different from the d_E-covariance field (d_E being the Euclidean metric). In order to compare the d_S-covariance and d_E-covariance fields visually, one could in principle use dimension reduction methods; however the interpretation of projections of the d_S-covariance may be problematic, as discussed in Section 3.1 of the main paper. An alternative way to represent the d-covariance variations is to consider a single location of interest and plot the distances

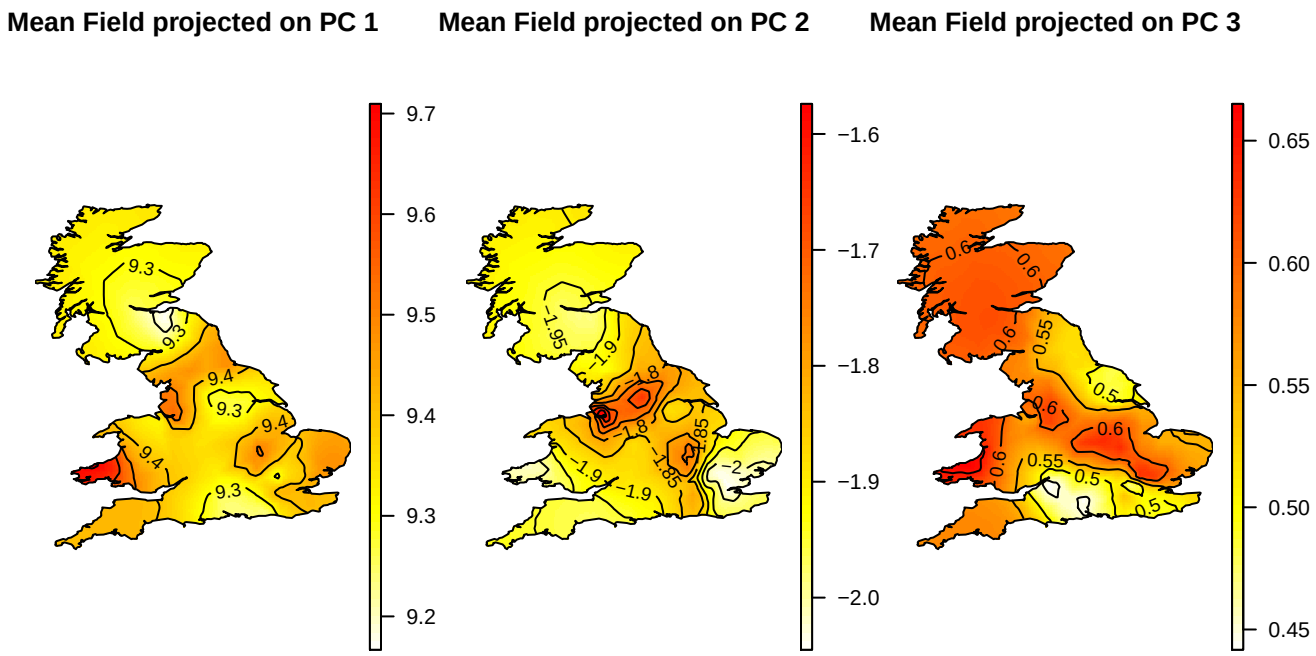


Figure S9: Projection onto PCs 1,2,3 of the mean obtained from data simulated under the global model (S5.1)

between the d-covariance at the location of interest, and the d-covariances at all other locations of the map. This produces $2D$ surfaces that reflect which parts of the country are more similar or dissimilar to the location of interest with respect to d-covariance. Figure S10 shows an example of these distance surface for the square-root and the Euclidean metric, where the distance between d-covariances has been computed using the d_S metric in both cases (averaged over the length of the sound), and the distances have been renormalized to the interval $[0, 1]$ to allow for fair comparison of the plots. The tuning parameters are $h = 1, k = 32$ nearest locations. Notice that the level curves are different. In particular, the level curve 0.6 for the square-root map goes down to Bristol, whereas it goes down to Dorset in the Euclidean metric map. The level curve 0.8 is also very different between the two maps. These differences can be attributed to the swelling effect of the Euclidean metric (Arsigny et al. 2007).

References

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2007), 'Geometric means in a novel vector space structure on symmetric positive-definite matrices', *SIAM Journal on Matrix Analysis and Applications* **29**(1), 328–347.

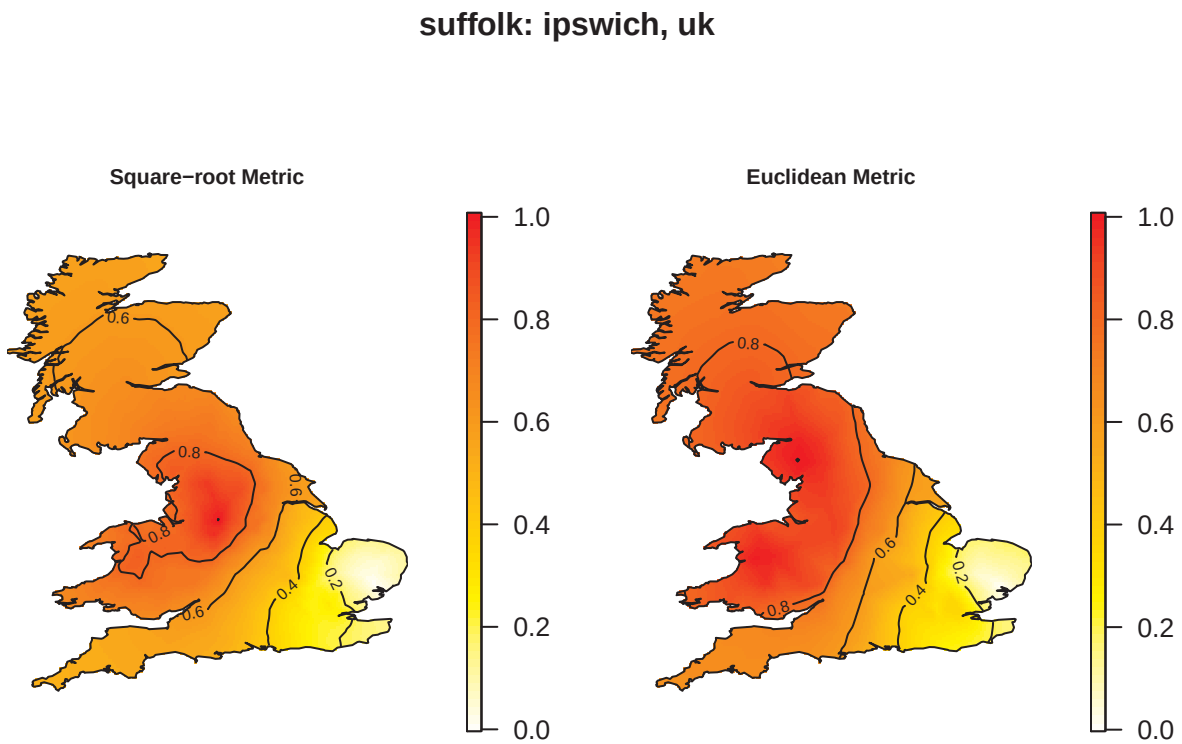


Figure S10: Left: color map with contours of the pairwise distances between the d_S-covariance at Ipswich (Suffolk), and the d_S-covariance at other locations. Right: same color map, but for the d_E-covariance. The tuning parameters are $h = 1, k = 32$ nearest locations. The scale of each map has been renormalized so that the value 1 is the maximal pairwise distance (under the metric d_S) in the d-covariance field, respectively for each metric.

Axler, S. (2015), *Linear algebra done right*, 3rd ed. edn, Cham: Springer.

van der Loo, M. P. (2014), 'The stringdist Package for Approximate String Matching', *R Journal* **6**(1), 111–122.