جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Extended Excess Hazard Models for Spatially Dependent Survival Data

André V. Ribeiro Amaral [†]

andre.ribeiroamaral@kaust.edu.sa

King Abdullah University of Science and Technology (KAUST)
Geospatial Statistics and Health Surveillance Research Group

[†] Joint work with Francisco Javier Rubio, Manuela Quaresma, Francisco J. Rodríguez-Cortés, and Paula Moraga.

Introduction

Goal: Under the "relative survival framework," we aim at proposing a general parametric class of frailty models and inference tools that account for spatial dependence (and investigate geographical inequalities in England)[1].

When modeling survival data, we may work under three different frameworks

1. **Overall survival framework:** no distinction is made among the possible causes of death. This is not useful when comparing populations.
2. **Cause-specific framework:** used when the cause of death is known. For certain diseases and locations, this information may not be available.
3. **Relative survival framework:** used when we can decompose the hazard function into "hazard associated with other causes" and "hazard associated with some specific disease." Useful when comparing populations.

[1] Amaral, A. V. R., Rubio, F. J., Quaresma, M., Rodríguez-Cortés, F. J., and Moraga, P. (2023). Extended Excess Hazard Models for Spatially Dependent Survival Data. arXiv. https://arxiv.org/abs/2302.09392.

Introduction

Let $T_1, \cdots, T_n$ represent the time-to-death for n individuals, such that $T_i \overset{\text{i.i.d.}}{\sim} G$, $\forall i$.

In that case, we are interested in the following functions

- Survival function: $S(t) = \mathbb{P}(T > t) = 1 - F(t)$, such that $F(t)$ is the CDF.
- Hazard function:

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt},$$

  such that $f(t)$ is the PDF. $h(t)$ can be interpreted as the rate of death at t. The hazard functions is also used to characterize the distribution of T.
- Cumulative Hazard function: $H(t) = \int_0^t h(s)ds$, such that $S(t) = \exp\{-H(t)\}$.

Relative survival framework

Under the relative survival framework, and assuming we are analyzing cancer-diagnosed patients,

$$h(t; \mathbf{x}) = h_{\mathrm{O}}(\mathrm{age} + t) + h_{\mathrm{E}}(t; \mathbf{x}), \ t \geq 0,$$

where h_{O} is the hazard associated with other causes of death, and h_{E} is associated with cancer. h_{E} is also known as excess hazard. Also, "age" is the patient's age when diagnosed, and $\mathbf{x}$ is a vector of risk factors.

As h_{O} is usually not known in practice, we approximate it by the population hazard h_{P} (estimated by life-tables). In that way,

$$h(t; \mathbf{x}, \mathbf{z}) = h_{\mathrm{P}}(\mathrm{age} + t; \mathbf{z}) + h_{\mathrm{E}}(t; \mathbf{x}), \ t \geq 0,$$

such that $\mathbf{z} \subseteq \mathbf{x}$ is a vector of patient characteristics (e.g., age, sex, race, etc.).

Relative survival framework

When comparing populations using models fitted under the relative survival framework, people usually refer to the net survival, that is

$$S_{\mathrm{N}}(t;\mathbf{x}) = \exp\left\{-\int_0^t h_{\mathrm{E}}(s;\mathbf{x})ds\right\},$$

where h_{E} is the excess hazard. Alternatively, we can report the population net survival

$$S_{\mathrm{N}}(t) = \frac{1}{m}\sum_{i=1}^{m} S_{\mathrm{N}}(t;\mathbf{x}_i),$$

where $\{\mathbf{x}_i\}_{i=1}^m$ represents the set of covariates associated to the (sub-)population of interest.

Excess hazard model

Let t_{ij} be the observed survival times in the i-th region for the j-th patient, and δ_{ij} be the censoring indicators. Then, the model for $h_{\mathrm{E}}(\cdot)$ is

$$h_{\mathrm{E}}(t; \mathbf{x}_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tilde{u}_i, u_i) = h_0(t \exp\{\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\alpha} + \tilde{u}_i\} \mid \boldsymbol{\theta}) \exp\{\mathbf{s}_{ij}^\top \boldsymbol{\gamma} + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + u_i\}, \qquad (1)$$

where $h_0(\cdot \mid \boldsymbol{\theta})$ is the baseline function (defined through a parametric distribution), $\boldsymbol{\theta}$ represents the corresponding parameters, $\tilde{\mathbf{x}}_{ij} \subseteq \mathbf{x}_{ij}$, such that $\tilde{\mathbf{x}}_{ij} \in \mathbb{R}^{\tilde{p}}$ and $\mathbf{x}_{ij} \in \mathbb{R}^p$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{\tilde{p}})^\top$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$. Also, $\mathbf{s}_{ij} = ({s_{ij1}}^\top, \cdots, {s_{ijq}}^\top)^\top \in \mathbb{R}^q$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)^\top$, where $q = \sum_{\ell=1}^k q_\ell$, such that q_ℓ is the dimension of $s_{ij\ell}$, and $s_{ij\ell}$ is the spline expansion of a (continuous) covariate $x_{ij\ell}$.

The random effects will follow $\tilde{\mathbf{u}} = (\tilde{u}_i, \cdots, \tilde{u}_r)^\top \sim \tilde{G}$ and $\mathbf{u} = (u_1, \cdots, u_r)^\top \sim G$, such that $\tilde{G}$ and G are spatially (in)dependent models (e.g., IID, ICAR, BYM2, etc.).

Notice that Model (1)—name it "Relative Survival Spatial General Hazard" (RS-SGH) model—generalizes the following approaches

Table 1: All possible model extensions based on the Relative Survival Spatial General Hazard (RS-SGH) approach.

Name	Description	Name	Description
RS-SGH-I	$\tilde{\mathbf{u}} = \mathbf{0}$	RS-GH	$\tilde{\mathbf{u}} = \mathbf{u} = \mathbf{0}$
RS-SGH-II	$\tilde{\mathbf{u}} = \mathbf{u}$	RS-PH	$\tilde{\mathbf{u}} = \mathbf{u} = \mathbf{0}, \boldsymbol{\alpha} = \mathbf{0}$
RS-MEPH	$\tilde{\mathbf{u}} = \mathbf{0}, \boldsymbol{\alpha} = \mathbf{0}$	RS-AFT	$\tilde{\mathbf{u}} = \mathbf{u} = \mathbf{0}, \boldsymbol{\alpha} = \boldsymbol{\beta}$
RS-MEAFT	$\tilde{\mathbf{u}} = \mathbf{u}, \boldsymbol{\alpha} = \boldsymbol{\beta}$	RS-AH	$\tilde{\mathbf{u}} = \mathbf{u} = \mathbf{0}, \boldsymbol{\beta} = \mathbf{0}$

Inference

Let $\mathcal{D} = \{(t_{ij}, \delta_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij});\ i = 1, \ldots, r, \text{ and } j = 1, \ldots, n_i\}$ be the observed data. In that case, the likelihood function for the vector of unknown parameters can be written as proportional to

$$\prod_{i=1}^{r} \prod_{j=1}^{n_i} \left\{ h_{\mathrm{P}}(\mathrm{age}_{ij} + t_{ij}; \mathbf{z}_{ij}) + h_{\mathrm{E}}(t_{ij}; \mathbf{x}_{ij} \mid \boldsymbol{\xi}, \tilde{u}_i, u_i) \right\}^{\delta_{ij}} \exp\{-H_{\mathrm{E}}(t_{ij}; \mathbf{x}_{ij} \mid \boldsymbol{\xi}, \tilde{u}_i, u_i)\}, \quad (2)$$

where $h_{\mathrm{P}}(\mathrm{age}_{ij} + t_{ij}; \mathbf{z}_{ij})$ is obtained from the life tables.

From Equation (2), notice that the only term in the likelihood function that distinguishes an overall survival model from a relative survival model is $h_{\mathrm{P}}(\mathrm{age}_{ij} + t_{ij}; \mathbf{z}_{ij})$; thus, by setting it to zero, we could also retrieve the overall survival framework.

Implementation

As Model (1) is fully parametric, we will set the baseline $h_0(\cdot)$ structure as

1. Log-normal,
2. Log-logistic,
3. Power Generalized Weibull (not Weibull, due to non-identifiability issues),
4. Gamma (it is tricky to fit, since it depends on a special function), and
5. Generalized Gamma (same issues as for the Gamma).

And the random effects $\tilde{\mathbf{u}} = (\tilde{u}_i, \cdots, \tilde{u}_r)^\top$ and $\mathbf{u} = (u_1, \cdots, u_r)^\top$ will be set as

1. IID model; i.e., $\tilde{u}_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2_{\tilde{u}})$, $\forall i$ (same for $\tilde{\mathbf{u}}$),
2. Intrinsic Conditional Autoregressive (ICAR) model, or
3. Besag-York-Mollié (BYM2) model.

Implementation

Table 2: Implemented models based on the baseline distributions and the random effects structures.

#	Dist.	Model	R.E.	#	Dist.	Model	R.E.	#	Dist.	Model	R.E.	#	Dist.	Model	R.E.	#	Dist.	Model	R.E.
01	LN	RS-SGH	BYM2	20	LL	RS-SGH	BYM2	39	PGW	RS-SGH	BYM2	58	GAM	RS-SGH	BYM2	77	GG	RS-SGH	BYM2
02	LN	RS-SGH	ICAR	21	LL	RS-SGH	ICAR	40	PGW	RS-SGH	ICAR	59	GAM	RS-SGH	ICAR	78	GG	RS-SGH	ICAR
03	LN	RS-SGH	IID	22	LL	RS-SGH	IID	41	PGW	RS-SGH	IID	60	GAM	RS-SGH	IID	79	GG	RS-SGH	IID
04	LN	RS-SGH-I	BYM2	23	LL	RS-SGH-I	BYM2	42	PGW	RS-SGH-I	BYM2	61	GAM	RS-SGH-I	BYM2	80	GG	RS-SGH-I	BYM2
05	LN	RS-SGH-I	ICAR	24	LL	RS-SGH-I	ICAR	43	PGW	RS-SGH-I	ICAR	62	GAM	RS-SGH-I	ICAR	81	GG	RS-SGH-I	ICAR
06	LN	RS-SGH-I	IID	25	LL	RS-SGH-I	IID	44	PGW	RS-SGH-I	IID	63	GAM	RS-SGH-I	IID	82	GG	RS-SGH-I	IID
07	LN	RS-SGH-II	BYM2	26	LL	RS-SGH-II	BYM2	45	PGW	RS-SGH-II	BYM2	64	GAM	RS-SGH-II	BYM2	83	GG	RS-SGH-II	BYM2
08	LN	RS-SGH-II	ICAR	27	LL	RS-SGH-II	ICAR	46	PGW	RS-SGH-II	ICAR	65	GAM	RS-SGH-II	ICAR	84	GG	RS-SGH-II	ICAR
09	LN	RS-SGH-II	IID	28	LL	RS-SGH-II	IID	47	PGW	RS-SGH-II	IID	66	GAM	RS-SGH-II	IID	85	GG	RS-SGH-II	IID
10	LN	RS-MEPH	BYM2	29	LL	RS-MEPH	BYM2	48	PGW	RS-MEPH	BYM2	67	GAM	RS-MEPH	BYM2	86	GG	RS-MEPH	BYM2
11	LN	RS-MEPH	ICAR	30	LL	RS-MEPH	ICAR	49	PGW	RS-MEPH	ICAR	68	GAM	RS-MEPH	ICAR	87	GG	RS-MEPH	ICAR
12	LN	RS-MEPG	IID	31	LL	RS-MEPH	IID	50	PGW	RS-MEPH	IID	69	GAM	RS-MEPG	IID	88	GG	RS-MEPG	IID
13	LN	RS-MEAFT	BYM2	32	LL	RS-MEAFT	BYM2	51	PGW	RS-MEAFT	BYM2	70	GAM	RS-MEAFT	BYM2	89	GG	RS-MEAFT	BYM2
14	LN	RS-MEAFT	ICAR	33	LL	RS-MEAFT	ICAR	52	PGW	RS-MEAFT	ICAR	71	GAM	RS-MEAFT	ICAR	90	GG	RS-MEAFT	ICAR
15	LN	RS-MEAFT	IID	34	LL	RS-MEAFT	IID	53	PGW	RS-MEAFT	IID	72	GAM	RS-MEAFT	IID	91	GG	RS-MEAFT	IID
16	LN	RS-GH	—	35	LL	RS-GH	—	54	PGW	RS-GH	—	73	GAM	RS-GH	—	92	GG	RS-GH	—
17	LN	RS-PH	—	36	LL	RS-PH	—	55	PGW	RS-PH	—	74	GAM	RS-PH	—	93	GG	RS-PH	—
18	LN	RS-AFT	—	37	LL	RS-AFT	—	56	PGW	RS-AFT	—	75	GAM	RS-AFT	—	94	GG	RS-AFT	—
19	LN	RS-AH	—	38	LL	RS-AH	—	57	PGW	RS-AH	—	76	GAM	RS-AH	—	95	GG	RS-AH	—

Code implementation is available on `https://github.com/avramaral/relative_survival`.

Model selection

To compare the fitted models, we will use a leave-one-out cross validation (LOO CV) procedure. That is, we will use the likelihood evaluated at the parameters' posterior samples as a goodness-of-fit measure.

Under the Bayesian framework, the LOO estimate of out-of-sample predictive fit will be computed as

$$\text{elpd}_{\text{LOO}} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log \left[\pi(t_{ij} \mid \mathbf{t}_{-ij}) \right],$$

where $\pi(t_{ij} \mid \mathbf{t}_{-ij})$ is the LOO predictive density given $\mathbf{t}_{-ij}$, such that $\mathbf{t}_{-ij}$ corresponds to the vector of all observed time points, except t_{ij}.

As a remark, the larger $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$ (the estimate for elpd_{LOO}), the better.

Applications

We will analyze a data set that contains survival information about male and female patients diagnosed with colon cancer between 2015 and 2016 in England.

We will present two case studies

1. Patients in all England with spatial structure defined in two different manners
   1.1 Based on the administrative boundaries given by the Government Office Regions.
   1.2 Based on the health boundaries determined by the Cancer Alliances.
2. Patients diagnosed with colon cancer in London. However, their locations (according to the London Clinical Commissioning Groups), will be defined based on
   2.1 The patients' areas of residence.
   2.2 The areas where patients receive treatment.

The population hazard term $h_{\mathrm{P}}(\mathrm{age}_{ij} + t; \mathbf{z}_{ij})$ was determined based on the life tables for England defined for the corresponding calendar year, and stratified by age, sex, deprivation level, and region of residence.

Case study 01

For our first analysis, we fit Model (1) for $10,936$ male patients and $9,586$ female patients diagnosed in England with colon cancer in 2016.

The linear predictor terms are defined by

$$\text{Time-level}: \ \text{age}_{ij}\alpha + \tilde{u}_i$$

$$\text{Hazard-level}: \ \text{age}_{ij}\beta_1 + \sum_{k=2}^{K} \mathbb{1}_{\text{stage}_{ij}(k)}\beta_k + \text{deprivation}_{ij}\beta_{(K+1)} + u_i,$$

where $\mathbb{1}_{\text{stage}_{ij}(k)}$, for $2 \leq k \leq 4$, is an indicator function for individuals who belong to the k-th cancer tumour stage ("1" being *early stage* and "4" *late stage*)

We fit the following models: RS-SGH LL ICAR, RS-SGH LL BYM2, RS-SGH LN ICAR, RS-SGH LN BYM2, RS-SGH PGW ICAR, and RS-SGH PGW BYM2, as per the notation introduced in Table 2.

Next, the best model is selected according to the $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$ criterion (RS-SGH LN BYM2).

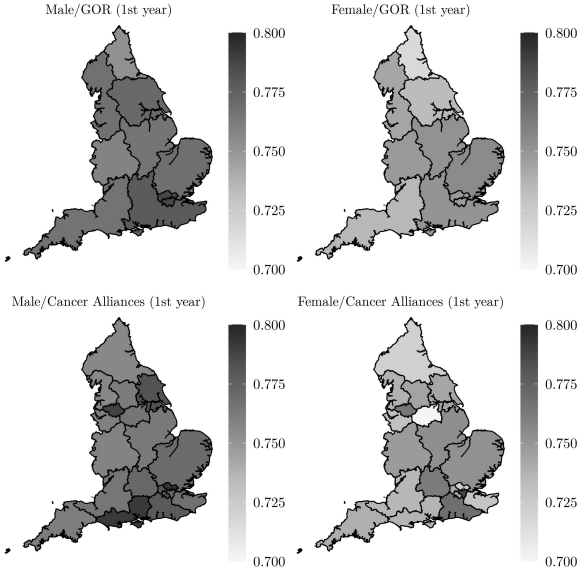

Figure 1: Net survival point estimate for $t = 1$ with model RS-SGH LN BYM2 for all classes.

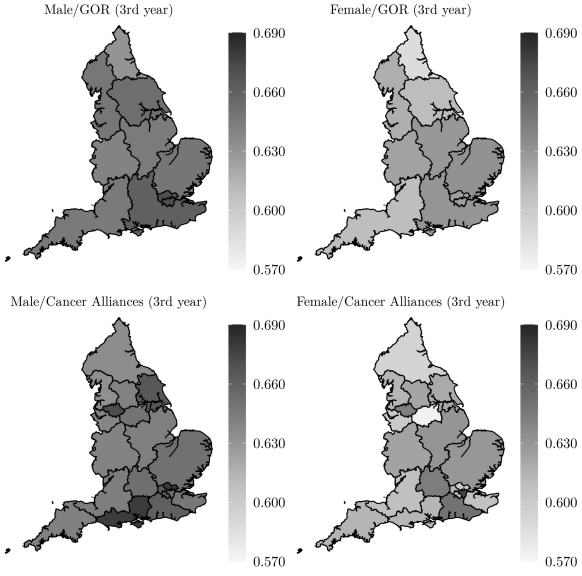

Figure 2: Net survival point estimate for $t = 3$ with model RS-SGH LN BYM2 for all classes.

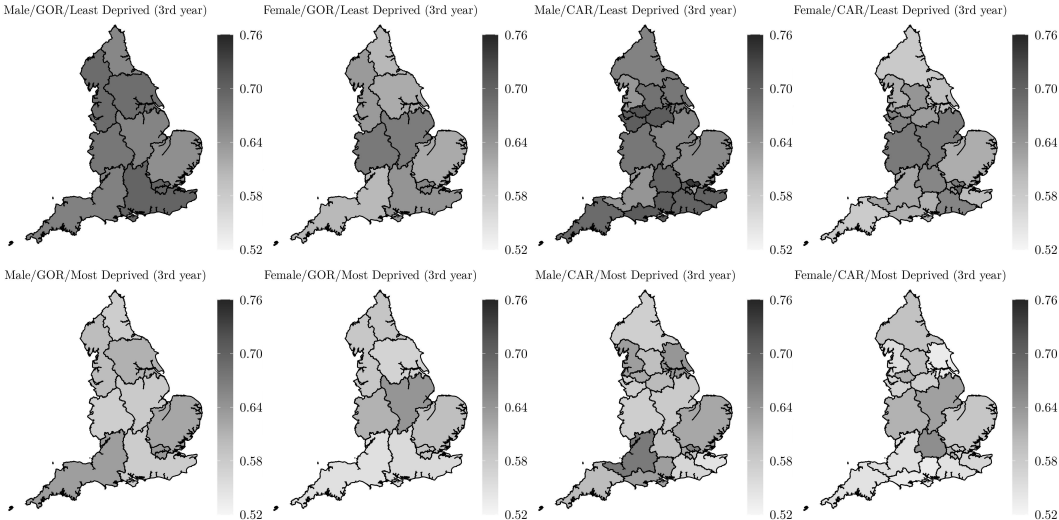

Figure 3: "Deprivation level" ("1" being *least deprived* (top row) and "5" *most deprived* (bottom row)) stratified net survival point estimate for $t = 3$ with model RS-SGH LN BYM2 for all classes.

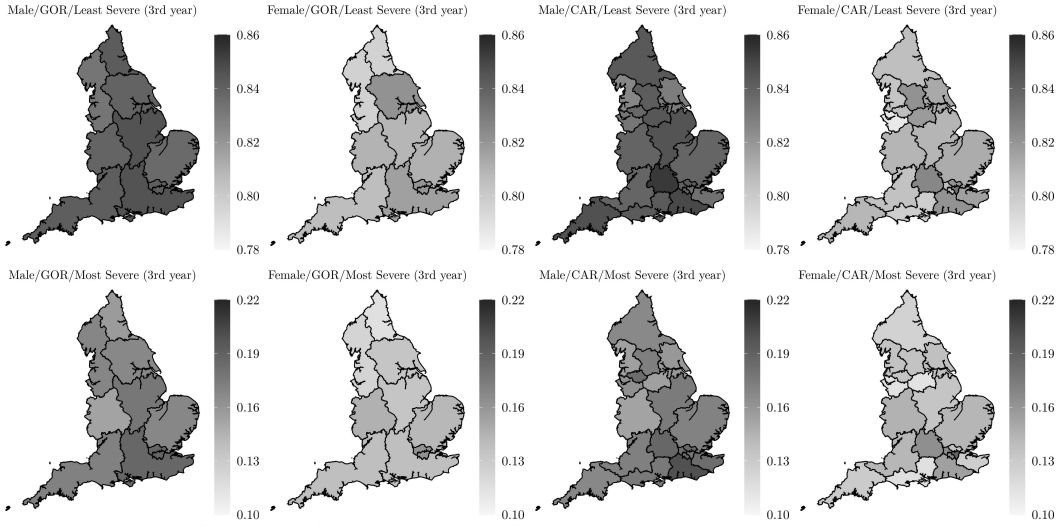

Figure 4: "Cancer stage" ("1, 2, and 3" being *least severe* (top row) and "4" *most severe* (bottom row)) stratified net survival point estimate for $t = 3$ with model RS-SGH LN BYM2 for all classes.

Case Study 02

For our second analysis, we fit Model (1) for $1,342$ male patients and $1,135$ female patients diagnosed in London with colon cancer in 2015 and 2016.

Our main goal is to investigate whether there seem to have differences in the net survival due to the patients' commuting habits when getting treated.

For that reason, we only account for $K = 3$ levels for cancer stage (patients with *stage 4*-cancer are likely not to receive treatment for cancer removal; instead, the treatment is focused on improving their quality of life).

For *area of residence*, we fit models RS-SGH LL ICAR, RS-SGH LL BYM2, RS-SGH LN ICAR, RS-SGH LN BYM2, RS-SGH PGW ICAR, and RS-SGH PGW BYM2. However, for *area of treatment*, we fit models RS-SGH LL IID, RS-SGH LN IID, and RS-SGH PGW IID.

Next, as before, the best model is selected according to the $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$.

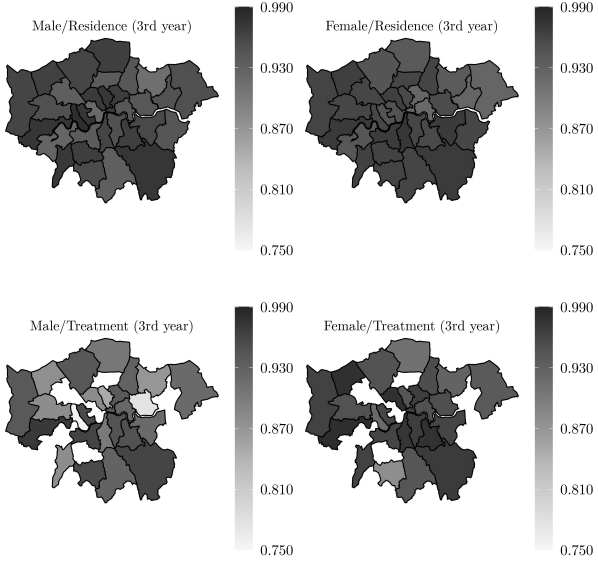

Figure 5: Net survival point estimate for $t = 3$ with models RS-SGH PGW BYM2 and LL IID.

Discussion

We proposed the Relative Survival Spatial General Hazard (RS-SGH) class of frailty models.

This work also contains other minor contributios, such as

1. Prior distribution recommendations for the model parameters and hyperparameters.
2. Guidelines about the sample size, baseline hazard distribution misspecification, and censoring rate when fitting models of this kind.

In this regard, based on a simulation study, we concluded that

1. Sample size and the censoring rate were shown to be the most important factors to control.
   1.1 Minimum sample size of 500–1000 patients estimate well the net survival curves.
   1.2 High censoring rates (e.g., 50%) with not large sample sizes (e.g., 200–500 patients) produced biased estimates (specially for 3-parameter distributions).
2. Misspecification of $h_0(\cdot)$, if we have enough non-censored data and a model that can capture the true hazard shape, had little negative impact.

Thank you!