道路交通碰撞数据时空模型的贝叶斯推断

Nicola Hewett[a]、Andrew Golightly[b]、Lee Fawcett[a] 和 Neil Thorpe[c]

[a]纽卡斯尔大学数学、统计与物理学院，泰恩河畔纽卡斯尔，NE1 7RU，英国[b]杜伦大学数学科学系，斯托克顿路，达勒姆，DH1 3LE，英国[c]纽卡斯尔大学工程学院，泰恩河畔纽卡斯尔，NE1 7RU，英国

arXiv:2302.00342v1 [stat.AP] 1 Feb 2023

摘要

提高道路安全至关重要，全球道路交通事故死亡人数仍高得令人无法接受；据估计，每年有135万人丧生（WHO，2020年）。目前，处理碰撞事故热点的做法几乎总是被动的：当某个预先设定的观测期内碰撞次数超过阈值后，才采取措施（例如安装道路安全摄像头）。然而，近年来已有方法被开发出来，用于预测未来各时段潜在热点的碰撞事故计数，以便更主动地处理道路安全热点。动态线性模型为预测碰撞事故提供了灵活的框架，从而能够采取此类主动措施。本文展示了如何使用这类模型来捕捉多个地点随时间变化的碰撞率中的季节变异和空间依赖。该模型能够对数据完整的地点以及部分数据缺失的地点进行样本内和样本外预测。我们采用美国佛罗里达州北部8个交通管理区的碰撞率数据说明这一方法，并发现该模型能够很好地描述潜在过程，且具有合理的预测准确度。

关键词：动态线性模型（DLM）；贝叶斯推断；前向滤波后向抽样器；马尔可夫链蒙特卡罗

1 引言

每年约有 130 万人的生命因道路交通事故而过早结束。另有 2,000 万至 5,000 万人遭受非致命伤害，其中许多人因伤致残（世界卫生组织，2020年）。使用碰撞事故计数可能会引发零膨胀问题，尤其是在较短的时间范围内。通过使用各区域的发生率，我们可以减少数据集中的零值；剔除这些零值后，还可以将数据视为连续数据，这在建立易于处理的模型时具有数学上的便利性。大多数道路交通数据都是按时间顺序记录的，各观测值之间通常存在依赖关系。因此，必须通过时间序列模型（例如状态空间模型）在模型中考虑这些依赖关系。状态空间模型在道路安全分析中的应用相对较新，也并不常见，但这类模型有利于预测。

状态空间模型可用于对非平稳、结构变化和不规则模式下的单变量或多变量时间序列进行建模（例如，参见 Harvey，1990年；West 和 Harrison，2006年）。时间序列分析通常从构建能够解释时间依赖性的模型开始，例如通过自相关、趋势或季节性来解释。

在时间序列场景中使用状态空间模型，可以量化观测过程以及任何未被直接观测到的动态变量中的不确定性。因此，预测会考虑这些不同的不确定性来源；在贝叶斯范式下进行推断时，还会考虑额外的参数不确定性。本文始终聚焦于一类特定的状态空间模型，其中观测方程和系统方程均由潜在过程的线性函数构成。这类模型称为动态线性模型（DLM；例如参见 Petris 等人，2009年；West 和 Harrison，2006年），与非线性模型相比具有多项实用优势。尤其是，这类模型具有易于处理的观测数据似然函数，从而能够高效地进行推断和预测。

Gamerman 和 Migon（1993年）列出了一系列用于状态演化、平滑以及沿层级各阶段进行滤波的分层动态线性模型（动态线性模型）。尽管迄今为止，状态空间模型，尤其是动态线性模型，在道路安全领域很少得到应用（例如，参见 Buddhavarapu，2015年；Fei 等人，2011年），但它们已广泛应用于环境领域。例如，Lai 等人（2020年）采用时空模型，对英格兰东北部五个位置的温度和湿度测量值所构成的传感器输出进行预测。该信号通过耦合动态线性模型描述，其空间效应由高斯过程（高斯过程）指定。Shaddick 和 Wakefield（2002年）在排放数据领域提出了一种相关方法。

我们的贡献是构建一个针对多个区域碰撞率的联合时空模型。在单区域层面使用动态线性模型，并通过一个谐波纳入季节性，其振幅和相位参数随时间变化。随后，我们在系统方程中加入空间高斯过程，以考虑邻近位置之间的空间依赖，从而平滑相对于潜在时间过程的空间偏差。所得模型既可对观测完整的位置进行样本内和样本外预测，也可对部分数据缺失的位置进行预测。我们采用贝叶斯方法推断动态和静态模型组成部分，并利用观测数据似然的易处理性；该似然可通过前向滤波器高效计算（例如，参见 Carter 和 Kohn，1994年；Fr¨uhwirth-Schnatter，1994年）。我们将这一推断方案应用于一项真实数据研究，研究对象为美国佛罗里达州北部固定交通管理区内的月碰撞率数据。我们先分别评估每个区域的季节性分量是否由随时间变化的参数控制这一假设，再考虑所有区域的联合模型。

本文其余部分安排如下。第 2 节简要介绍数据。第 3 节介绍单区域和联合区域的动态线性模型结构。第 4 节概述贝叶斯推断方案的细节，然后在第 5 节考察真实数据应用。第 6 节总结全文。

2 数据

我们考察的是美国佛罗里达州北部固定交通管理区内的月碰撞率数据。共有 8 个区域在多个地点跟踪并记录了碰撞事故计数。随后，根据每个区域内各地点每月的平均碰撞次数，计算出该区域的碰撞率。每个区域有 115 个月的观测值，最新观测值来自 2014年4月。图1展示了不同区域随时间变化的多条数据流。所有区域的月碰撞率都呈现出以 12 个月为周期的正弦模式。月碰撞率的直方图表明，高斯观测模型或许能够充分描述观测过程。通过散点图，我们发现年份 t 至 $t+1$ 之间的某些月份存在明显的时间依赖性，因此无法采用将“月份”作为固定效应的简单模型。此外，地理位置越近的区域，相关性越强（见图2）。

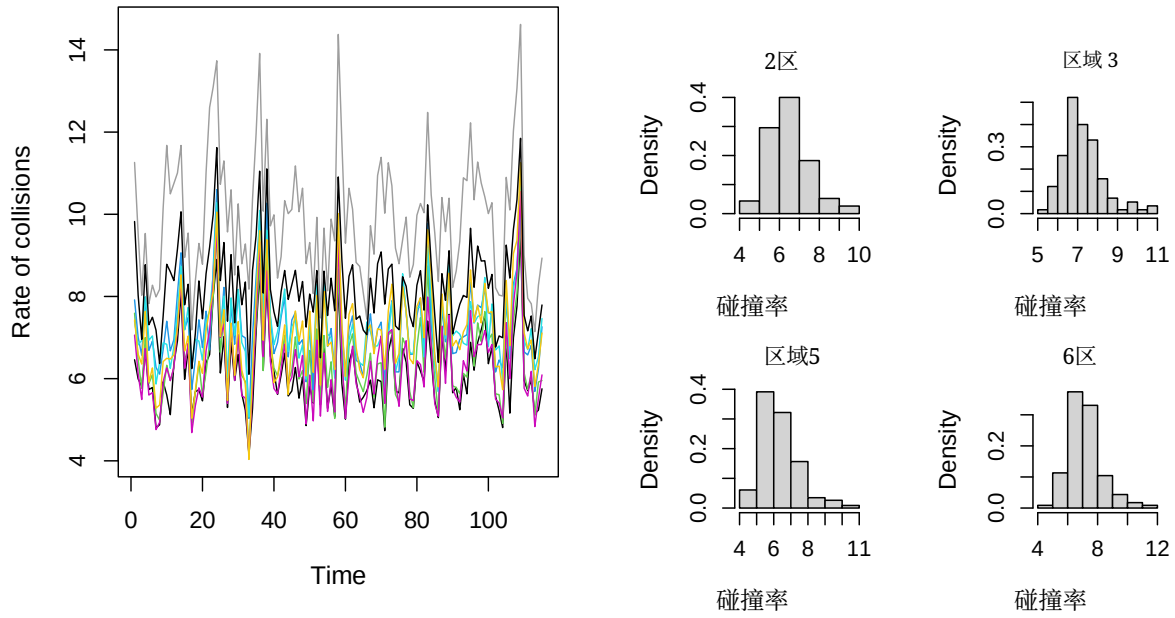

图1：左：8个区域的月碰撞率时间序列图。右：区域2、3、5、6的碰撞率直方图。

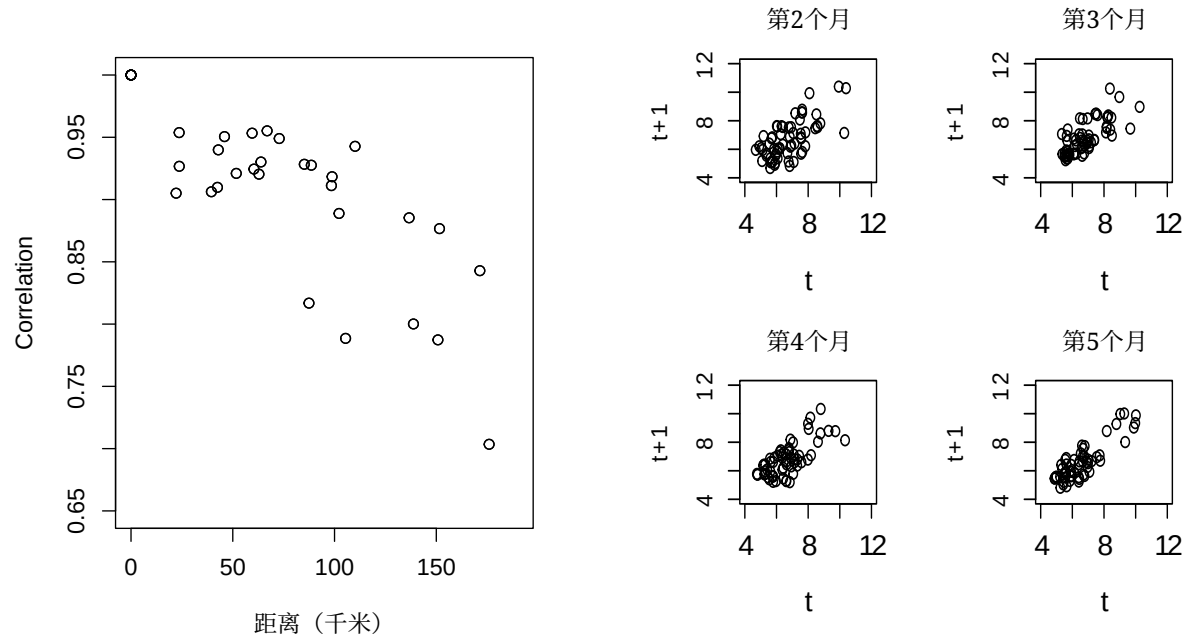

图2：左：8个区域之间的相关性与区域间距离（千米）的关系。右：所有区域中连续年份第2、3、4、5个月观测值之间的时间依赖性。

3 动态线性模型（DLM）

状态空间模型建立在相对简单的一阶马尔可夫链依赖结构之上（也就是说，关于某个状态 θ_{t_i} 的信息由链中截至时间 t_{i-1} 的所有先前值所携带，与仅由 $\theta_{t_{i-1}}$ 所携带的信息相同）。它们由两个主要部分组成：观测数据（x_{t_1}，……，x_{t_n}）和未观测/潜在状态（θ_{t_0}，……，θ_{t_n}）。图 3 展示了一个简单单变量状态空间的演变，其中连续取值的潜在状态过程$\{\theta_{t_0}, \theta_{t_1}, \ldots, \theta_{t_{i-1}}, \theta_{t_i}, \ldots\}$遵循一阶马尔可夫链，其转移密度为$\pi(\theta_{t_i}|\theta_{t_{i-1}})$。连续取值的观测过程 $\{x_{t_1}, x_{t_2}, \ldots, x_{t_{i-1}}, x_{t_i}, \ldots\}$ 通过密度 $\pi(x_{t_i}|\theta_{t_i})$与任意时刻 t_i 的潜在状态过程相关联；这里假定给定潜在状态后，观测数据条件独立。可观测过程 $\{X_{t_i}\}$取决于其底层不可观测的潜在状态过程 $\{\theta_{t_i}\}$，我们可以合理地假设，观测值 X_{t_i} 仅取决于进行测量时系统的状态 θ_{t_i}。接下来需要明确 X_{t_i} 与 θ_{t_i}之间，以及 θ_{t_i} 与$\theta_{t_{i-1}}$之间的关系。在这两种情况下，我们都采用线性关系，并进一步假设状态部分和观测部分中的误差相互独立且服从正态分布。这种结构构成了动态线性模型（DLM），其方程如下：

$$\begin{aligned} \text{Observation Equation}: \quad & X_{t_i} = F_{t_i}\theta_{t_i} + \nu_{t_i} \\ \text{System Equation}: \quad & \theta_{t_i} = G_{t_i}\theta_{t_{i-1}} + \omega_{t_i} \end{aligned}$$

此处，X_{t_i} 是标量，θ_{t_i} 是 $p \times 1$ 向量，F_{t_i} 是 $1 \times p$ 向量，G_{t_i} 是 $p \times p$ 矩阵，而 ν_{t_i} ~N$(0, V_{t_i})$和 ω_{t_i} ~N$(0, W_{t_i})$是相互独立的白噪声过程，其方差矩阵 V_t和 W_t已知，通常假设为常数。假设初始潜在状态服从高斯分布，则有

$$\begin{aligned} \theta_0 &\sim \mathrm{N}(m_0, C_0) \\ \theta_{t_i}|\theta_{t_{i-1}} &\sim \mathrm{N}(G_{t_i}\theta_{t_{i-1}}, W_{t_i}) \\ X_{t_i}|\theta_{t_i} &\sim \mathrm{N}(F_{t_i}\theta_{t_i}, V_{t_i}) \end{aligned}$$

其中超参数 m_0 和 C_0经过了适当选择。接下来，我们先考虑适用于单个区域数据的动态线性模型，再考虑覆盖所有区域的联合模型。

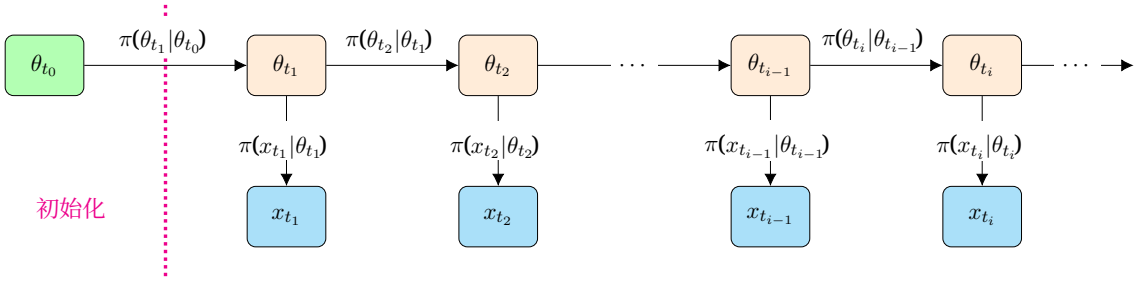

图 3：展示状态空间模型依赖结构的有向无环图。

3.1 分区特定模型

第 2 节所述的数据集呈现出季节性：在所有区域中，碰撞率在一年内都表现出明显的正弦模式。因此，为了在动态线性模型中考虑这一点，我们加入一个谐波。请注意，也可以通过在系统方程中加入多个谐波来考虑季节性（例如，参见 Petris 等人，2009年）；不过，我们发现，采用单个谐波并允许振幅和相位随时间变化，是一种简约的建模方法。

首先考虑单个位置。我们假设方差矩阵 V 和 W 为常数，且数据的观测时间间隔不规则： t_1， t_2，……， t_n。观测方程为

$$X_{t_i} = F_{t_i}\theta_{t_i} + \nu_{t_i}, \quad \nu_{t_i} \overset{indep}{\sim} \mathrm{N}(0, V), \tag{1}$$

其中 $\theta_{t_i} = (\theta_{1,\ t_i},\ \theta_{2,\ t_i},\ \theta_{3,\ t_i})^T$，观测矩阵定义为

$$F_{t_i} = \left(\sin\left(\frac{2\pi i}{P_x}\right), \cos\left(\frac{2\pi i}{P_x}\right), 1\right)$$

其中 P_x 表示一个完整周期对应的时间（季节性数据中为 $P_x = 12$ ）。注意，观测方程可写为

$$X_{t_i} = \tilde{\theta}_{1,t_i}\cos\left(\frac{2\pi i}{P_x} - \tilde{\theta}_{2,t_i}\right) + \theta_{3,t_i} + \nu_{t_i} \tag{2}$$

其中，式（1）和式（2）中的动态参数通过以下关系相联系：

$$\tilde{\theta}_{1,t_i} = \sqrt{\theta^2_{1,t_i} + \theta^2_{2,t_i}}, \qquad \tilde{\theta}_{2,t_i} = \tan^{-1}\left(\frac{\theta_{1,t_i}}{\theta_{2,t_i}}\right). \tag{3}$$

我们通过将系统方程设为以下形式，对这些动态参数施加一定的平滑性约束：

$$\theta_{t_i} = \theta_{t_{i-1}} + k_{t_i}\omega_{t_i}, \quad \omega_{t_i} \overset{indep}{\sim} \mathrm{N}(0, W)$$

该方程经过进一步调整，以适用于在时间网格上间隔不规则的测量值。也就是说，我们在状态方程的方差中加入一个系数 k_{t_i}，使得 $k^2_{t_i} = t_i - t_{i-1}$。因此，正弦形式的动态线性模型通过单个谐波捕捉季节性，其振幅 $\tilde{\theta}_{1,\ t_i}$ 和相位 $\theta_{2,\ t_i}$ 随两个经过变换的独立随机游走过程而变化。

3.2 区域联合模型

现在我们考虑一种月碰撞率模型，该模型既能捕捉季节性，也能捕捉相邻区域之间的相关性。令 $X_{t_i} = (X^1_{t_i},\ \ldots\ldots,\ X^{n_z}_{t_i})^T$ 表示时间 t_i 时的月碰撞率集合，其中 $X^j_{t_i}$ 对应区域 j，且 $j = 1,\ \ldots\ldots,\ n_z$。在第 5.1 节中，我们发现振幅和相位对于每个区域而言都可能是恒定的。因此，为简化符号，以下固定 $\theta_{1,\ t_i} = \theta_1$ 和 $\theta_{2,\ t_i} = \theta_2$。区域 j 的模型为

$$X^j_{t_i} = \theta^j_1 \sin\frac{\pi t_i}{6} + \theta^j_2 \cos\frac{\pi t_i}{6} + \theta^j_{3,t_i} + \nu^j_{t_i}, \quad \nu^j_{t_i} \overset{indep}{\sim} N(0, V^j),$$
$$\theta^j_{3,t_i} = \theta^j_{3,t_{i-1}} + k_{t_i}\omega^j_{t_i} + p^j_{t_i}, \quad \omega^j_{t_i} \overset{indep}{\sim} N(0, W^j),$$

为引入相邻区域之间的相关性，我们将项 $p^j_{t_i}$ 作为空间平滑误差过程 $p_{t_i} = (p^1_{t_i},\ \ldots\ldots,\ p^{n_z}_{t_i})^T$ 的一个组成部分。我们使用相互独立（沿 i）的零均值高斯过程来建模 $\{p_{t_i},\ t_i = 1, \ldots, n\}$，因此

$$p_{t_i} \overset{indep}{\sim} GP\{0, f_3(\cdot;\eta_3)\}.$$

我们通过采用平方指数核作为协方差函数来施加平滑性。因此，位置 j 和 j' 处空间误差之间的协方差为

$$f_3(d_{jj'};\eta_3) = \mathrm{Cov}(\theta^j_{3,t_i}, \theta^{j'}_{3,t_i}) = \sigma^2_3 \exp(-\phi_3 d_{jj'}), \tag{4}$$

其中 $\eta_3 = (\sigma_3, \phi_3)^T$ 用于参数化核；请注意，随着地点 j 与 j' ($d_{jj'}$) 之间的距离增大，ϕ_3 决定了相关性的衰减比率（Banerjee等（2014））。同样，

我们为 $\theta_1 = (\theta_1^1, \ldots, \theta_1^{n_z})^T$ 和 $\theta_2 = (\theta_2^1, \ldots, \theta_2^{n_z})^T$ 采用高斯过程先验，使得 $\theta_1 \sim GP(m_1(\cdot)\ f_1(\cdot;\ \eta_1))$ 和 $\theta_2 \sim GP(m_2(\cdot)\ f_2(\cdot;\ \eta_2))$，其中 f_1 和 f_2 的定义与公式（4）类似，并根据需要添加 $m_1(\cdot)$ 和 $m_2(\cdot)$ 作为均值函数。因此，完整的空间动态线性模型（涵盖所有位置）为

$$
\begin{aligned}
X_{t_i} &= F_{t_i}\theta_{t_i} + \nu_{t_i}, \quad \nu_{t_i} \overset{indep}{\sim} \mathrm{N}(0, \mathrm{diag}\{V^1, \ldots, V^{n_z}\}), \\
\theta_{3,t_i} &= \theta_{3,t_{i-1}} + k_{t_i}\omega_{t_i}, \quad \omega_{t_i} \overset{indep}{\sim} \mathrm{N}(0, \mathrm{diag}\{W^1, \ldots, W^{n_z}\} + K_3),
\end{aligned}
$$

其中 $F_{t_i} = \mathrm{diag}(F_{t_i}^1, \ldots, F_{t_i}^{n_z})$, $\theta_{3,t_i} = (\theta_{3,t_i}^1, \ldots, \theta_3^{n_z},_{t_i})^T$, $\theta_{t_i} = (\theta_1^1, \theta_2^1, \theta_{3,t_i}^1, \ldots, \theta_1^{n_z}, \theta_2^{n_z}, \theta_3^{n_z},_{t_i})^T$和$K_3$是一个 $n_z \times n_z$ 矩阵，其 $(i,j)th$ 元素为 $f_3(d_{ij}, \eta_3)$。

4 贝叶斯推断

为简化起见，假设我们有 n_z 个区域，每个区域有 n 个观测值。令 $V = (V^1, \ldots, V^{n_z})^T$和 $W = (W^1, W^2, \ldots, W^{n_z})^T$。此外，令 $\eta_3 = (\sigma_3, \phi_3)^T$表示控制 $f_3(\cdot)$ 的超参数，其中 $\eta_1 = (\sigma_1, \phi_1)^T$和 $\eta_2 = (\sigma_2, \phi_2)^T$分别表示控制 $f_1(\cdot)$ 和 $f_2(\cdot)$ 的超参数。令 $x^j = (x_{t_1}^j, \ldots, x_{t_n}^j)^T$表示地点 j 的碰撞率向量，因此 $x = (x^1, \ldots, x^{n_z})$ 表示所有区域的完整数据集。所有动态参数和静态参数的联合后验，与边际静态参数后验乘以动态过程 $\theta_3 = (\theta_{3,t_0}, \ldots, \theta_{3,t_n})$ 的条件后验成正比，使得

$$
\begin{aligned}
\pi(\theta_1, \theta_2, V, W, &\eta_1, \eta_2, \eta_3, \theta_3 | x) \propto \\
&\pi(\theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3 | x) \times \pi(\theta_3 | \theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3, x).
\end{aligned}
$$

令 ψ 表示所有固定模型参数。为从联合后验中模拟实现值，我们采用两步法:

1. 从边际后验 $\psi \sim \pi(\psi|x)$ 中模拟。
2. 从条件后验 $\theta_3 \sim \pi(\theta_3|\psi, x)$ 中模拟。

对于第 1 步，由于边际静态参数后验无法解析求解，我们采用马尔可夫链蒙特卡罗（例如，参见 Gilks 等人，1995年）。对于第 2 步，我们采用前向滤波后向抽样算法（例如，参见 West 和 Harrison，2006年），直接从 $\pi(\theta_3|\psi, x)$ 中抽样。详情如下。

4.1 基于模拟的推断

令 $\theta_{3,t_{0:n}} = (\theta_{3,t_0}, \theta_{3,t_1}, \ldots, \theta_{3,t_n})$ 表示截至时间 t_n 的潜在状态集合，并令 $x = x_{t_{1:n}} = (x_{t_1}, \ldots, x_{t_n})$ 表示观测数据。注意，$\theta_{3,t_i} = (\theta_{3,t_i}^1, \ldots, \theta_3^{n_z},_{t_i})^T$ 和 $x_{t_i} = (x_{t_i}^1, \ldots, x_{t_i}^{n_z})^T$。假设 ψ 的各组成项采用独立先验设定后，贝叶斯推断可按如下方式进行。对动态参数积分消去后，可得到边际后验:

$$
\begin{aligned}
\pi(\theta_1, \theta_2, &V, W, \eta_1, \eta_2, \eta_3 | x) \propto \\
&\pi(\theta_1|\eta_1)\pi(\theta_2|\eta_2) \left[\prod_{j=1}^{n_z} \pi(V^j)\pi(W^j)\right] \times \\
&\pi(\eta_1)\pi(\eta_2)\pi(\eta_3) \times \pi(x|\theta_1, \theta_2, \eta_3, V, W)
\end{aligned}
$$

其中，边际似然 $\pi(x|\theta_1,\ \theta_2,\ \eta_3,\ V,\ W)$ 的表达式为

$$
\pi(x|\theta_1, \theta_2, \eta_3, V, W) = \pi(x_{t_1}|\theta_1, \theta_2, \eta_3, V, W) \prod_{i=1}^{n-1} \pi(x_{t_{i+1}}|x_{t_{1:i}}, \theta_1, \theta_2, \eta_3, V, W) \tag{5}
$$

其组成项均可进行解析处理。此外，$\pi(\theta_1|\eta_1) = N(\theta_1;\ m_1,\ K_1)$ 和$\pi(\theta_2|\eta_2) = N(\theta_2;\ m_2,\ K_2)$ 是多元正态密度，$\pi(V^j)$ 和 $\pi(W^j)$ 是分别赋予 V^j和 W^j的先验密度，$\pi(\eta_1)$、$\pi(\eta_2)$ 和 $\pi(\eta_3)$ 是分别赋予 η_1、η_2 和 η_3的先验密度。

边际似然可通过前向滤波器高效计算。这里定义以下内容会有所帮助：

$$\tilde{X}_{t_i} \equiv X_{t_i} - \theta_1 \sin\frac{\pi t_i}{6} - \theta_2 \cos\frac{\pi t_i}{6} = \tilde{F}_{t_i}\theta_{3,t_i} + \nu_{t_i}, \quad \nu_{t_i} \overset{indep}{\sim} \mathrm{N}(0, \mathrm{diag}\{V^1, \ldots, V^{n_z}\}),$$

使得

$$\tilde{X}_{t_i}|\theta_{3,t_i} \sim N(\tilde{F}_{t_i}\theta_{3,t_i}, \mathrm{diag}\{V\}),$$

˜其中 F_{t_i} 是 $n_z \times n_z$ 单位矩阵，为简化后续记号将省略它。我们还记作

$$\theta_{3,t_i}|\theta_{3,t_{i-1}} \sim N(\theta_{3,t_{i-1}}, \tilde{W}_{t_i})$$

其中 $\tilde{W}_{t_i} = k_{t_i}^2$(diag$\{W^1, \ldots, W^{n_z}\} + K_3$)。

算法 1 前向滤波器

1. 初始分布：$\theta_{3,\ t_0} \sim N(m_0,\ C_0)$。存储 m_0和 C_0的值。2. 对于 t_i，$i = 1,$ . . . , n,

(a) t_i处的先验。根据系统方程，我们有

$$\theta_3|\tilde{x}_{t_{1:i-1}} \sim N(m_{t_{i-1}}, C_{t_{i-1}} + \tilde{W}_{t_i}).$$

˜存储 $R_{t_i} = C_{t_{i-1}} + W_{t_i}$.

(b) 一步预测。根据观测方程，我们有

$$\tilde{X}_{t_i}|\tilde{x}_{t_{1:i-1}} \sim N(m_{t_{i-1}},\ R_{t_i} + V).$$

存储边际似然贡献

$$\pi(\tilde{x}_{t_i}|\tilde{x}_{t_{1:i-1}}) = N(\tilde{x}_{t_i}; m_{t_{i-1}}, R_{t_i} + V).$$

(c) t_i处的后验：$\theta_{3,t_i}|\tilde{x}_{t_{1:i}} \sim N(m_{t_i}, C_{t_i})$，其中

$$m_{t_i} = m_{t_{i-1}} + R_{t_i}(R_{t_i} + V)^{-1}(\tilde{x}_{t_i} - m_{t_{i-1}}),$$
$$C_{t_i} = R_{t_i} - A_{t_i}Q_{t_i}A_{t_i}^T,$$

其中 $A_{t_i} = R_{t_i}Q_{t_i}^{-1}$和 $Q_{t_i} = R_{t_i} + V$。存储 m_{t_i}和 C_{t_i}的值。

算法 1 给出了前向滤波器的步骤。可以看到，公式（5）中的各项是通过前向遍历得到的：

$$\pi(\tilde{x}_{t_i}|\tilde{x}_{t_{1:i-1}}, \theta_1, \theta_2, \eta_3, V, W) = \mathrm{N}(\tilde{x}_{t_i}; m_{t_{i-1}}, R_{t_i} + V),$$

˜其中 $R_{t_i} = C_{t_{i-1}} + W_{t_i}$和 $m_{t_{i-1}}$，$C_{t_{i-1}}$均以递归方式更新；有关更多细节，请参阅 Petris 等（2009）（另见 Carter 和 Kohn，1994，Fr¨uhwirth-Schnatter，1994，West 和 Harrison，2006）。

尽管边际似然是可处理的，但后验通常无法写成闭式形式。因此，我们使用 Metropolis-Hastings 算法从 $\pi(\psi|\tilde{x})$ 中生成抽样值；参见算法2。

算法2 马尔可夫链蒙特卡罗方案

1 使用 $\psi^{(0)}$ 初始化链。设置 $r=1$。2 提议 $\psi^* \sim q(\psi^*|\psi^{(r-1)})$。3 计算所提议移动的接受概率 $\alpha(\psi^*|\psi^{(r-1)})$，其中

$$\begin{aligned}\alpha(\psi^*|\psi^{(r-1)}) &= \min\left\{1, A(\psi^*|\psi^{(r-1)})\right\} \\ &= \min\left\{1, \frac{\pi(\psi^*|\tilde{x}_{1:n})q(\psi^{(r-1)}|\psi^*)}{\pi(\psi^{(r-1)}|\tilde{x}_{1:n})q(\psi^*|\psi^{(r-1)})}\right\}\end{aligned}$$

4 以概率 $\alpha(\psi^*|\psi^{(r-1)})$，设定 $\psi^{(r)}=\psi^*$；否则设定 $\psi^{(r)}=\psi^{(r-1)}$。5 设定 $r := r+1$。返回步骤2。

仍需注意，给定 $\psi^{(1)},\ldots,\psi^{(N)}$ 的抽样结果后，我们可以对 $\theta_3^{(r)} \sim \pi(\theta_3|\psi,x)$、 $r=1,\ldots,N$ 进行采样。注意到以下因式分解即可实现这一点:

$$\pi(\theta_3|\psi,x) = \pi(\theta_{3,t_n}|\psi,x_{t_{1:n}})\prod_{i=0}^{n-1}\pi(\theta_{3,t_i}|\theta_{3,t_{i+1}},\psi,x_{t_{1:i}})$$

其中各组成密度均易于处理，并可通过后向采样算法递归地进行抽样。关键步骤见算法3。

算法3 后向采样器

3. 对 $\theta_{3,n}|\tilde{x}_{t_{1:n}} \sim N(m_n, C_n)$. 进行采样。4. 对于 $t_i, i=n,\ldots,1$,

(a) 后向分布：$\theta_{3,\ t_i}|\theta_{3,\ t_{i+1}},\ \tilde{x}_{t_{1:i}} \sim N(h_{t_i},\ H_{t_i})$，其中

$$\begin{aligned}h_{t_i} &= m_{t_i} + C_{t_i}(C_{t_i}+\tilde{W}_{t_{i+1}})^{-1}(\theta_{3,t_{i+1}} - m_{t_i}), \\ H_{t_i} &= C_{t_i} - C_{t_i}^T(C_{t_i}+\tilde{W}_{t_{i+1}})^{-1}C_{t_i}.\end{aligned}$$

(b) 抽样 $\theta_{3,\ t_i}|\theta_{3,\ t_{i+1}},\ \tilde{x}_{t_{1:i}} \sim N(h_{t_i},\ H_{t_i})$。

缺失数据

缺失观测值很常见，也就是说，在时间 t_i，可能只有关于 X_t~部分组成成分的观测值可用。为在模型中考虑这一点，我们令 $X_{t_i}^o$ 表示时间 t_i的观测速率。观测模型可写为

$$\tilde{X}_{t_i}^o = P_{t_i}\tilde{X}_{t_i} \tag{6}$$

其中， $n_{obs}\times n_z$关联矩阵 P_{t_i}决定了在时间 t_i观测哪些组成成分（Lai 等人（2020））。例如，假设我们有来自 5 个区域的数据，并且在时间 t_i缺失第二个和第三区域的数据，则关联矩阵为

$$P_{t_i} = \begin{pmatrix}1&0&0&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}.$$

前向滤波器和后向采样器可以直接修改，以适用于此情~ ~形。简而言之，在算法 1 中，每处 F_{t_i} 都替换为 $P_{t_i}F_{t_i}$，每处 V 都替换为$P_{t_i}V$。

4.2 样本内预测密度

为了评估模型拟合，我们考虑样本内预测密度。样本内预测密度定义为

$$\pi(\hat{x}_{t_{1:n}}|x_{t_{1:n}}) = \int\int \pi(\hat{x}_{t_{1:n}}|\theta_{3,t_{1:n}},\psi)\pi(\theta_{3,t_{1:n}},\psi|x_{t_{1:n}})d\theta_{3,t_{1:n}}d\psi$$

其中

$$\pi(\theta_{3,t_{1:n}},\psi|x_{t_{1:n}}) = \pi(\theta_{3,t_{1:n}}|\psi,x_{t_{1:n}})\pi(\psi|x_{t_{1:n}}).$$

尽管样本内预测密度难以计算，但很容易从 $\pi(\theta_{3,t_{1:n}},\psi|x_{t_{1:n}})$ 中抽取样本，因此可以通过蒙特卡罗方法得到 $\pi(\hat{x}_{t_{1:n}}|x_{t_{1:n}})$。给定抽取的样本 $(\psi^{(r)},\theta_3^{(r)},_{t_{1:n}})$ $r=1,\ldots,N$，我们可以模拟

$$\hat{X}_{t_i}^{(r),j}|\theta_{t_i}^{(r),j},\psi^{(r),j} \sim N(F_{t_i}\theta_{t_i}^{(r),j},V^{(r),j}), \quad r=1,\ldots,N,\ i=1,\ldots,n,\ j=1,\ldots,n_z, \tag{7}$$

其中 $\theta^{(r)},_{t_i}{}^{j} = (\theta^{(r)},_1{}^{j},\theta^{(r)},_2{}^{j},\theta^{(r)},_3{}_{t_i}{}^{j},)$ 表示 $\theta_{t_i}^j$ 的第 r 个样本，$\hat{}X^{(r)},_{t_i}{}^{j}$ 的定义类似。根据 (7) 得到的样本可以进行汇总（例如通过均值、上分位数和下分位数），并与观测数据进行比较。

4.3 k 步 ahead 预测

利用动态线性模型的线性高斯结构，可以得到系统和观测预测分布。一步系统预测密度为

$$\begin{aligned}\pi(\theta_{3,t_{n+1}}|x_{t_{1:n}}) &= \int\int \pi(\theta_{3,t_{n+1}}|\theta_{3,t_n},\psi,x_{t_{1:n}})\pi(\theta_{3,t_n}|\psi,x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\theta_{3,t_n}d\psi \\ &= \int \pi(\theta_{3,t_{n+1}}|\psi,x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\psi\end{aligned}$$

其中

$$\pi(\theta_{3,t_{n+1}}|\psi,x_{t_{1:n}}) = N(\theta_{3,t_{n+1}};m_{t_n},C_{t_n}+\tilde{W}_{t_{n+1}}).$$

类似地，一步观测预测密度可表示为

$$\pi(x_{t_{n+1}}|x_{t_{1:n}}) = \int \pi(x_{t_{n+1}}|\psi,x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\psi$$

其中

$$\pi(x_{t_{n+1}}|\psi,x_{t_{1:n}}) = N(x_{t_{n+1}};m_{t_n},C_{t_n}+\tilde{W}_{t_{n+1}}+V).$$

因此，给定来自 $\pi(\theta_{3,t_n}|\psi,x_{t_{1:n}})$ 的 N后验摘要 $(m_{t_n}^{(r)},C_{t_n}^{(r)})$ $r=1,...,N$，以及来自$\pi(\psi|x_{t_{1:n}})$ 的 $\psi^{(r)}$，一步状态和观测预测分布可通过蒙特卡罗抽样获得，具体做法是抽取

$$\begin{aligned}\theta_{3,t_{n+1}}^{(r)}|\psi^{(r)},x_{t_{1:n}} &\sim N(m_{t_n}^{(r)},C_{t_n}^{(r)}+\tilde{W}_{t_{n+1}}^{(r)}), \\ \tilde{X}_{t_{n+1}}^{(r)}|\psi^{(r)},x_{t_{1:n}} &\sim N(m_{t_n}^{(r)},C_{t_n}^{(r)}+\tilde{W}_{t_{n+1}}^{(r)}+V^{(r)}).\end{aligned}$$

然后，将项 $\theta_1^{(r)}\sin\frac{\pi t_{n+1}}{6}+\theta_2^{(r)}\cos\frac{\pi t_{n+1}}{6}$ 加到 $\tilde{X}_{t_{n+1}}^{(r)}$ 上，即可得到 $X_{t_{n+1}}^{(r)}$。对于一般的 k步超前预测，可将上述抽样替换为

$$\begin{aligned}\theta_{n+k}^{(r)}|\psi,x_{t_{1:n}} &\sim N\left\{m_{t_n}^{(r)},R_{t_{n+k}}^{(r)}\right\}, \\ \tilde{X}_{t_{n+k}}^{(r)}|\psi^{(r)},x_{t_{1:n}} &\sim N\left\{m_{t_n}^{(r)},R_{t_{n+k}}^{(r)}+V^{(r)}\right\},\end{aligned}$$

其中

$$R_{t_{n+k}}^{(r)} = C_{t_n}^{(r)}+\sum_{i=1}^{k}\tilde{W}_{t_{n+i}}^{(r)}.$$

5 应用

在下文中，如有需要，我们采用第 4 节中的马尔可夫链蒙特卡罗方案，并使用具有高斯创新的随机游走提议。我们有 $q(\psi^*|\psi) = N(\psi^*; \psi, \Sigma)$，其中创新矩阵 $\Sigma = \gamma Var(\psi|x)$，$Var(\psi|x)$ 由预运行获得，而 γ 的取值旨在使接受率约为 25%（Roberts 和 Rosenthal，2001年）。在马尔可夫链蒙特卡罗方案中，为方便数学处理，我们将使用精度，因此 $\tau_V = 1/V$，$\tau_W = 1/W$。此外，对于分量必须严格为正的参数向量（即 V，W，η），我们将在对数尺度上实施提议。

5.1 单区域分析

本节评估振幅和相位是否随时间变化这一假设。我们展示4区的结果，并指出其余区域也有类似发现（即振幅和相位很可能保持不变）。

对于单区域模型，$\psi = (\tau_V, \tau_{W_1}, \tau_{W_2}, \tau_{W_3})'$ 是精度参数向量。我们将 θ_{t_0} 的均值和方差分别设为 $m_0 = (1.5, 1.5, 6)$和 $C_0 = \text{diag}\{1.5, 1.5, 20\}$。对于 ψ的各个分量，我们通过τ_V，τ_{W_1}，τ_{W_2}，$\tau_{W_3} \sim Ga(0.1, 0.1)$采用无信息且相互独立的先验设定。马尔可夫链蒙特卡罗方案运行了22k次迭代，舍弃前2k次迭代作为预热期，剩余20k次迭代用于计算后验摘要。

边际 MH 方案给出了 ψ 各分量的估计边际后验密度，如图4所示，并叠加了其先验密度。对 ψ 样本进行了稀疏抽取，从参数边际后验分布中获得 1k 个（近似不相关的）抽样，记为 $\{\psi^{(r)}\}_{r=1}^{1000}$。随后，对每个 $\psi^{(r)}$ 执行前向滤波后向抽样算法，从样本内预测分布中获取动态参数向量 $\{\theta_{t_i}^{(r)}\}_{r=1}^{1000}$、$i = 1$、.....、$n$ 的样本。动态分量 $\theta_{1、t_i}$ 和 $\theta_{2、t_i}$ 的样本可通过式（3）进行转换，从各自的样本内预测密度中获得相位和振幅的抽样；有关从这些预测分布中获取样本的方法，详见第 4.2 节。图5通过均值和95%可信区间对这些分布进行了概述。考虑到振幅和相位的不确定性，我们可以得出结论：对于该区域而言，它们随时间保持恒定是合理的。对其余区域进行分析后，也可得出相同结论。这表明，动态参数 $\theta_{1、t_i}$ 和 $\theta_{2、t_i}$、$i = 1$、.....、n 可以合理地替换为静态参数 θ_1 和 θ_2。

我们通过比较观测数据与基于模型的样本内后验预测分布，以及基于模型的样本外预测分布，评估所提出模型在单区域中的有效性。对于后者，我们在拟合模型时留出了最后10个观测值。图6展示了观测过程的样本内预测分布，并以每个时间点计算得到的均值和95%可信区间进行概括。这表明该模型能够较好地刻画观测过程。同样，10步 ahead 预测分布也以各时间点的均值和95%可信区间进行概括。可以看出，该预测分布能够捕捉观测值所呈现的总体趋势。

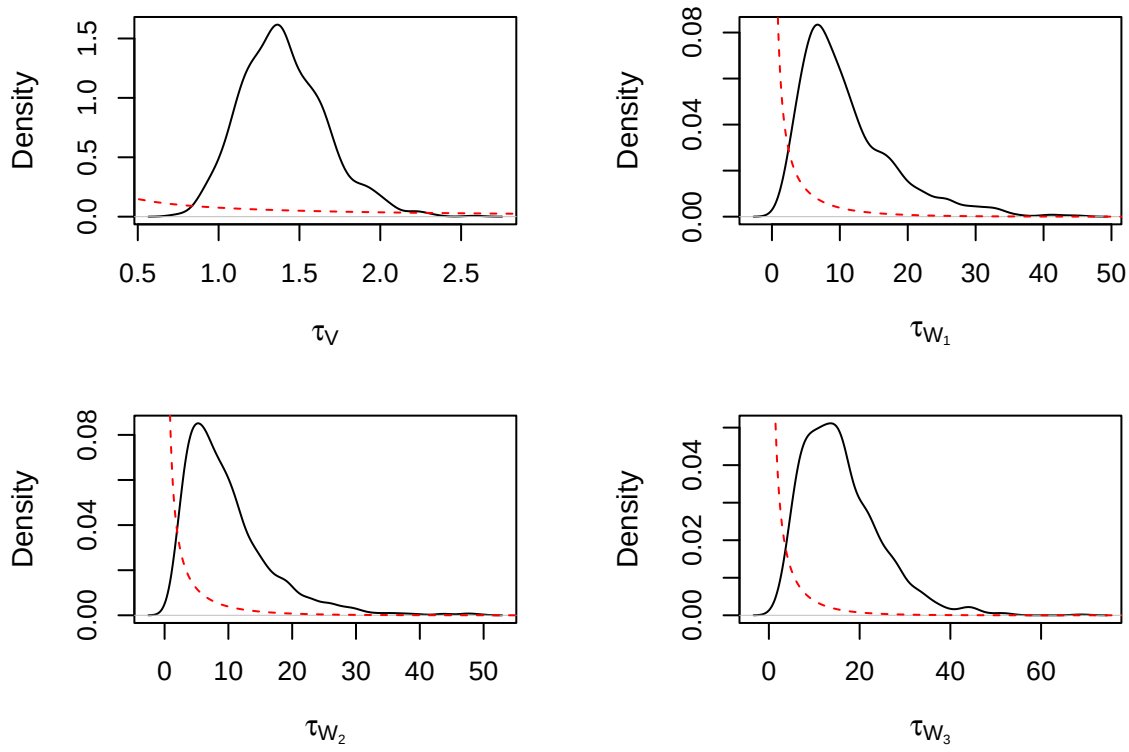

图4：分别基于 20k 次迭代、每 20 次抽取一次绘制的 τ_V、τ_{W_1}、τ_{W_2}、τ_{W_3} 密度图，并叠加红色先验密度。

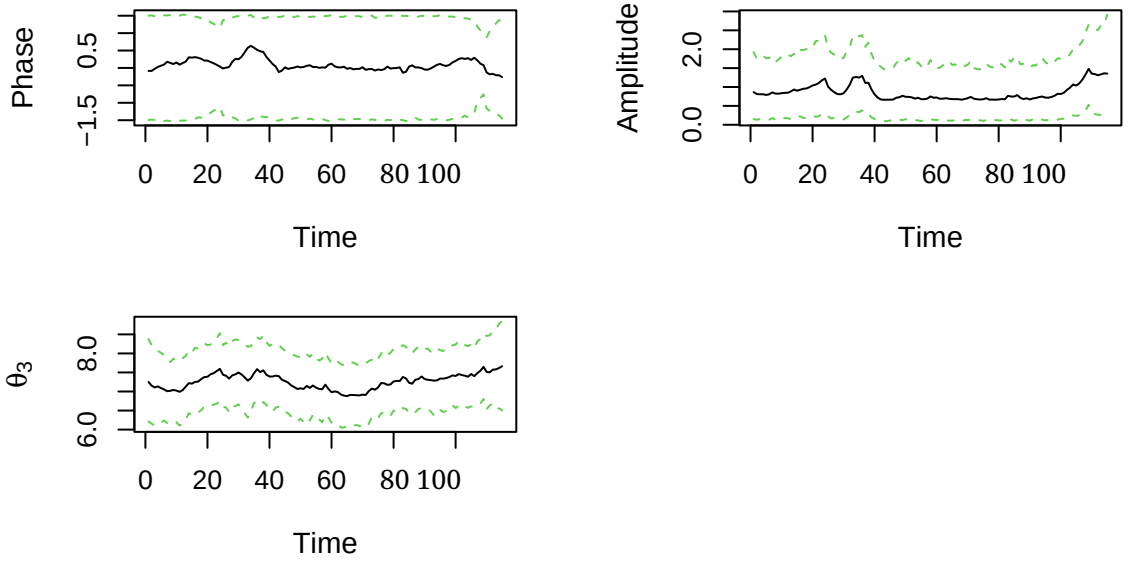

图5：时间 t_1 至 t_{11}期间，4区的相位、振幅和 θ_3、t_i 均值及 95%置信区间 $_5$.

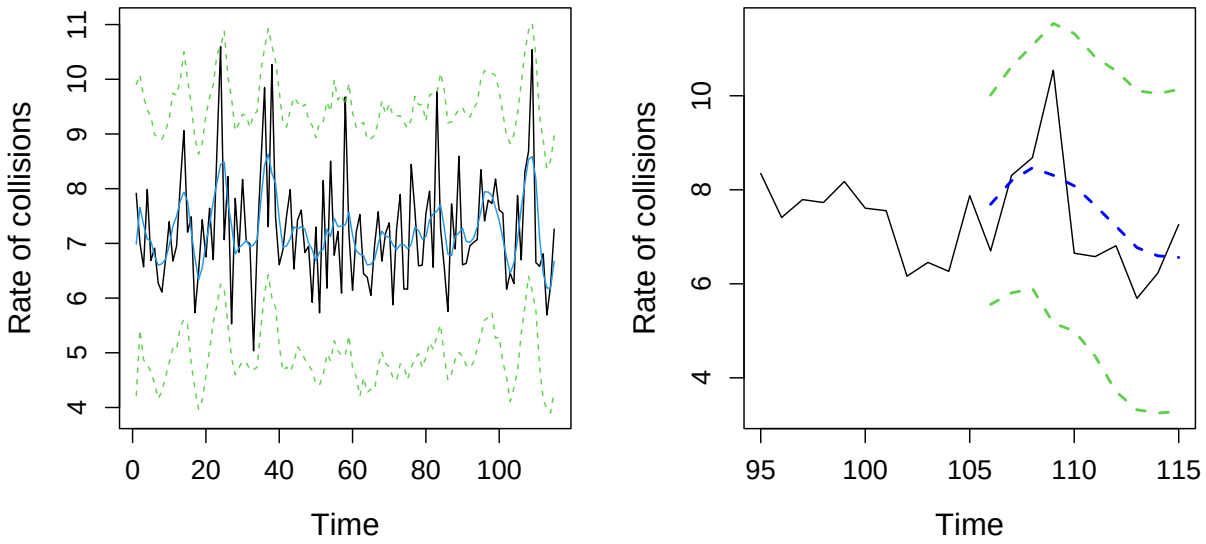

图6：左：4区观测数据（黑色）以及叠加的样本内预测——均值（蓝色）和95%可信区间（绿色）。右：观测数据（黑色）以及叠加的提前10步预测——均值（蓝色）和95%可信区间（绿色）。

5.2 联合区域分析

现在，我们考虑第3.2节详细介绍的涵盖所有区域的联合模型。我们的先验设定如下。

我们预期，相邻区域的振幅和相位应当相似。回想一下，$\theta_1 \sim GP(m_1(\cdot)\ f_1(\cdot;\ \eta_1))$、$\theta_2 \sim GP(m_2(\cdot)\ f_2(\cdot;\ \eta_2))$以及动态均值过程中的高斯过程分量均为 $p_{t_i} \overset{indep}{\sim} GP\{0, f_3(\cdot;\ \eta_3)\}$。我们将均值函数设为常数，因此$m_1(\cdot) = m_2(\cdot) = 1.5\mathbf{1}$，其中$\mathbf{1}$定义为由1构成的 $n_z \times 1$ 向量。我们有 $f_k(d_{jj'};\eta_k) = \sigma_k^2 \exp(-\phi_k d_{jj'})$，$k = 1$、2、3。我们取 $\log\ \sigma_k \overset{indep}{\sim} N(\log(0.1)\ 0.1)$，这代表了对区域内振幅方差和相位的较强先验信念。对于逆长度尺度的对数，我们取$\log\ \phi_k \sim N(\log(0.1)\ 0.1)$，对应的典型长度尺度约为10千米，反映区域间的典型距离。控制各区域观测方程的精度为$\tau_{Vj} \overset{indep}{\sim} Ga(0.1, 0.1)$，系统方差也类似，$\tau_{Wj} \overset{indep}{\sim} Ga(0.1, 0.1)$、$j = 1$、.....、$n_z$。最后，假设每个地点的初始值$\theta_{3t_0}^j$服从 $N(6, 20)$分布。

马尔可夫链蒙特卡罗方案运行了 1×10^6 次迭代；结果汇总于表1。图7展示了各区域振幅和相位后验密度均值及95%可信区间随经度的变化。相位似乎随着区域向东延伸而减小，振幅则增大，这表明存在空间依赖。图8展示了振幅和相位抽样值的平均结果所对应的一个正弦曲线周期，并比较了最东端与最西端区域及其95%可信区间。据此，我们预计东部区域全年碰撞率的波动会更加明显。此外，我们预计东部区域的碰撞率将在比西部区域早一个月时达到最高值（八月，而西部为九月）。

图9展示了区域2、4和8中观测值与样本内预测观测过程之差的摘要（均值和95%可信区间）。左侧图展示单区域分析所得的差值，右侧图展示联合分析所得的差值。显然，各时间-区域组合的平均差值都很小，并且

平均差值为零是合理的（95%可信区间包含零）。从左向右比较，可以看出，从单区域分析转为联合模型后，样本内预测有所改善；也就是说，通过高斯过程纳入的空间信息提高了预测精度。此外，我们还计算了每个时间点的均方根误差（RMSE，即观测值与预测值之差），并对各区域的所有时间点取该指标的平均值；结果见表2。可以看出，单区域分析的平均均方根误差约为联合模型的5倍，这进一步表明，与针对所有区域的联合模型相比，单区域分析的拟合效果较差。

图10展示了应用第4.3节的方法后，对区域2、4、6和8的提前10步预测。请注意，在运行推断方案之前，已从每个区域中移除最后10个观测值。图中显示，预测分布与数据一致，因为所有区域的预测分布都落在预测区间内。正如预期的那样，随着我们逐渐远离最后一次记录的观测值，不确定性会增加。

ψ	均值	95%置信区间	ψ	Mean	95%置信区间
V^1	0.034	(0.021, 0.052)	θ_1^4	0.251	(-0.045, 0.532)
V^2	0.025	(0.015, 0.039)	θ_1^5	0.226	(-0.065, 0.514)
V^3	0.059	(0.039, 0.084)	θ_1^6	0.249	(-0.039, 0.526)
V^4	0.037	(0.022, 0.058)	θ_1^7	-0.181	(-0.501, 0.133)
V^5	0.031	(0.018, 0.048)	θ_1^8	-0.014	(-0.308, 0.271)
V^6	0.041	(0.023, 0.066)	θ_2^1	0.585	(0.301, 0.877)
V^7	0.119	(0.059, 0.196)	θ_2^2	0.651	(0.367, 0.944)
V^8	0.045	(0.026, 0.071)	θ_2^3	0.566	(0.285, 0.856)
W^1	0.021	(0.011, 0.037)	θ_2^4	0.424	(0.144, 0.722)
W^2	0.024	(0.012, 0.041)	θ_2^5	0.809	(0.530, 1.098)
W^3	0.023	(0.011, 0.044)	θ_2^6	0.601	(0.311, 0.896)
W^4	0.025	(0.012, 0.044)	θ_2^7	1.264	(0.931, 1.587)
W^5	0.024	(0.013, 0.043)	θ_2^8	0.945	(0.660, 1.249)
W^6	0.034	(0.016, 0.061)	σ_1	1.688	(1.315, 2.309)
W^7	0.099	(0.031, 0.213)	σ_2	1.545	(1.201, 2.253)
W^8	0.029	(0.014, 0.055)	σ_3	1.352	(1.349, 1.355)
θ_1^1	0.357	(0.066, 0.642)	ϕ_1	1.527	(1.278, 1.917)
θ_1^2	0.213	(-0.077, 0.494)	ϕ_2	1.603	(1.387, 1.903)
θ_1^3	0.213	(-0.084, 0.499)	ϕ_3	1.103	(1.098, 1.107)

表1：通过马尔可夫链蒙特卡罗方案得到的参数边缘后验均值和基于分位数的95%可信区间。

	平均均方根误差	
Zone	单区域	联合区域
1	1.384	0.198
2	1.066	0.167
3	1.214	0.258
4	1.083	0.206
5	1.114	0.185
6	1.191	0.217
7	1.339	0.378
8	1.107	0.227

表2：单区域和联合区域分析中各区域所有时间点的均方根误差均值。

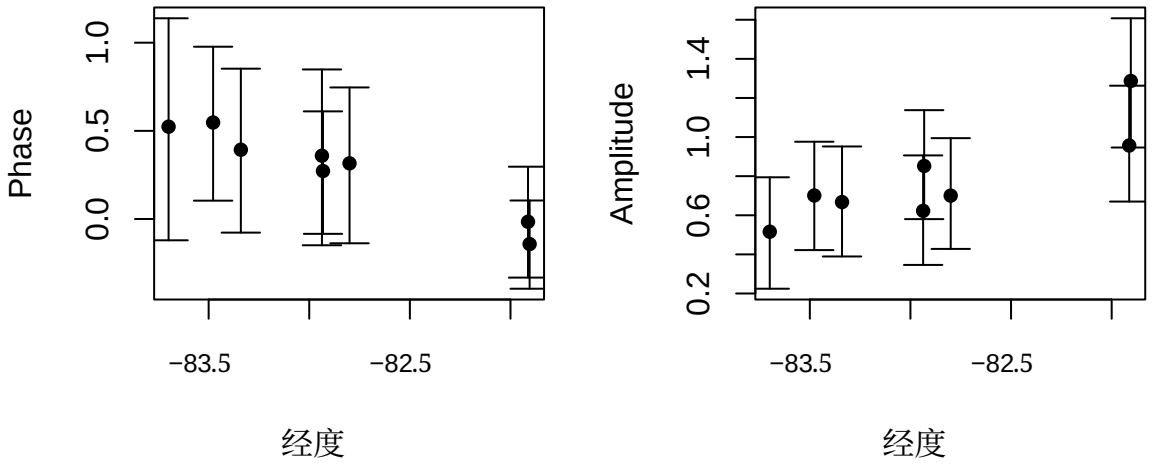

图7：均　　值振幅和相位的95%可信区间，按各区域的经度绘制　　e.

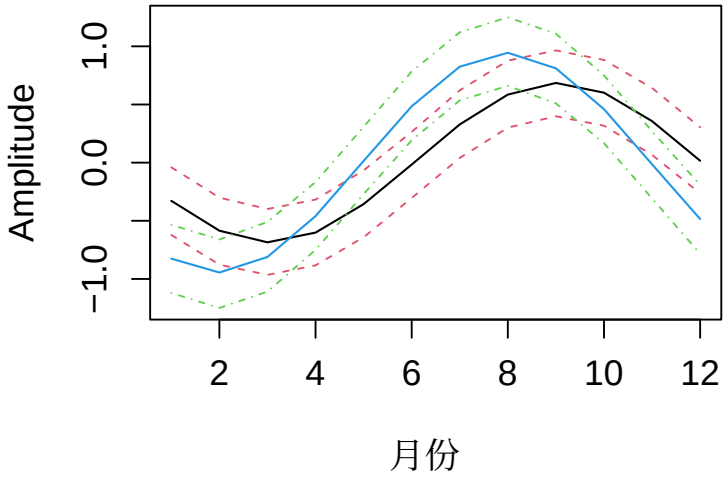

图8：最西侧区域（黑色）与最东侧区域（蓝色）的季节性分量均值及95%可信区间。

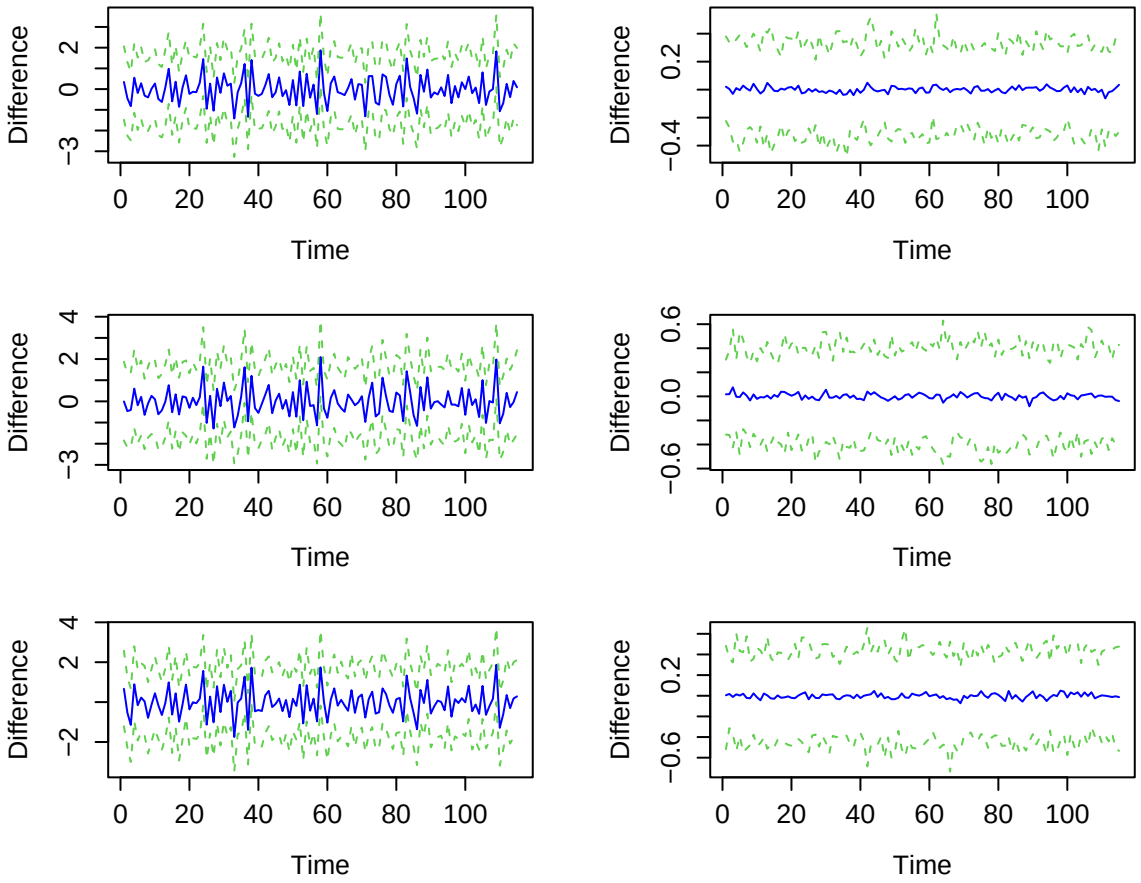

图9：随时间变化的样本内预测值与观测值之差的均值（蓝色）和95%可信区间（绿色）。每行分别展示区域2（上）、区域4（中）和区域8（下）的单区域分析（左）与联合区域分析（右）所得差值。

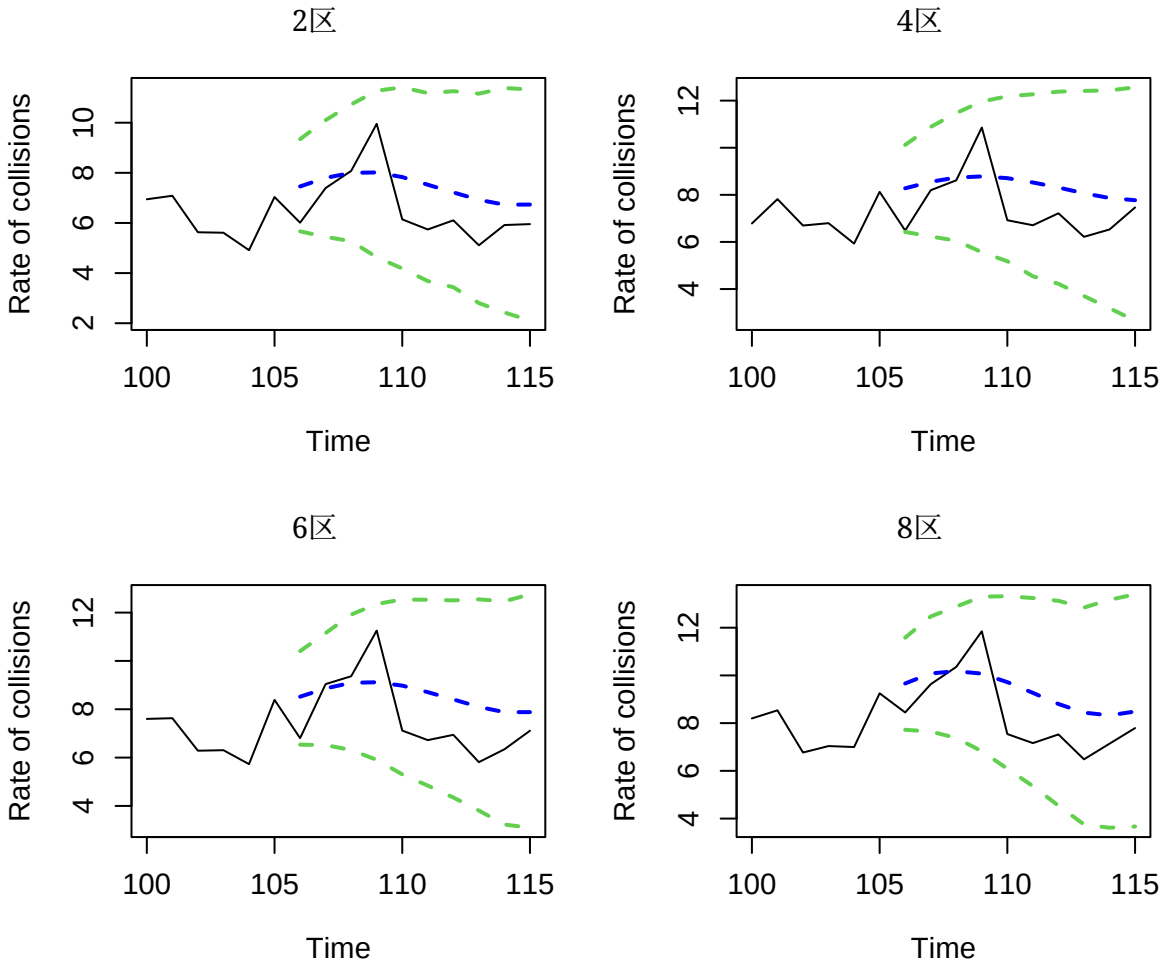

图10：佛罗里达州2、4、6和8区的碰撞率，并叠加了样本外提前10步预测——均值（蓝色）和95%可信区间（绿色）。

6 讨论与局限

我们开发了一种用于碰撞率的时空模型，该模型能够处理序列依赖、季节性以及邻近区域碰撞率之间的相关性。我们研究了一种动态线性模型（DLM），其观测方程采用单一谐波的形式，且振幅和相位随时间平滑变化，从而能够描述季节性和潜在的长期变化。通过在系统方程中加入高斯过程（GP）分量，该模型能够反映邻近区域之间的空间一致性。该模型可在贝叶斯范式下进行拟合，采用高效的两阶段马尔可夫链蒙特卡罗方法，针对参数、潜在时变谐波系数（振幅和相位）以及动态均值的联合后验进行推断。第一阶段使用随机游走 Metropolis 算法从参数边际后验分布中生成参数样本，并通过前向滤波器计算似然。第二阶段则以第一阶段抽取的静态参数为条件，使用后向采样器生成动态参数样本。关于这一前向滤波、后向采样（FFBS）方法的更多细节，参见 Petris 等（2009）（另见 Carter 和 Kohn，1994；Fr¨uhwirth-Schnatter，1994）。

我们将该方法应用于一个真实数据集，其中包含

美国佛罗里达州八个交通管理区连续115个月的碰撞率。对每个区域分别建立模型的探索性分析发现，相位和振幅可能保持不变。因此，我们将谐波分量视为静态参数，并使用允许相邻区域之间这些参数存在相关性的高斯过程先验，从而简化了所有区域的联合模型。我们使用样本内后验预测分布评估单区域模型和联合模型的有效性，结果表明两种模型的拟合均令人满意。此外，使用联合模型时，样本内预测得到了显著改善：预测的可信区间几乎缩小了十倍，观测值与预测值之间的均方根误差（RMSE）也降低了约5倍。

我们的分析表明，相位与经度、振幅与经度之间存在明显的空间模式。我们发现，所有区域的碰撞率最低点都出现在一年中较早的时候。模型还表明，西部区域的最低碰撞率出现在三月，而东部区域则出现在二月。东部的碰撞率峰值似乎出现在九月，西部则出现在八月。我们还预计，东部区域的碰撞率波动会更大。我们也关注能否预测未来几个月的碰撞率。基于模型的样本外预测分布表明，该模型能够捕捉月碰撞率中观测到的趋势和季节性，预测期最长约为一年。

我们的建模方法还可以通过多种方式加以改进。例如，通常会有与发生碰撞的特定位置相关的协变量信息，例如交通流量或平均速度。然而，按区域汇总此类数据既耗时，也并非总是简单易行。尽管如此，原则上，通过观测方程将协变量纳入动态线性模型框架并不困难，我们预计在这种情况下预测效果会有所改善。虽然本文未作探讨，我们的模型也可以用于预测那些没有观测值的区域的碰撞率。对于感兴趣但未观测到的区域，可以对系统方程中控制动态均值的拟合高斯过程分量，以及控制谐波的静态参数上的高斯过程先验进行插值；更多详情参见 Rasmussen 和 Williams（2005年）。

参考文献

Banerjee, S., Carlin, B. P. 和 Gelfand, A. E.（2014年）。空间数据的层次建模与分析。CRC出版社。Buddhavarapu, P. N. V. S. R.（2015年）。利用数据增广技术对空间与动态计数模型进行贝叶斯估计：在道路安全管理中的应用。博士论文。Carter, C. K. 和 Kohn, R.（1994年）。关于状态空间模型的吉布斯抽样。《生物统计学》，81(3)：541–553。Fei, X.、Lu, C.-C. 和 Liu, K.（2011年）。用于实时短期高速公路旅行时间预测的贝叶斯动态线性模型方法。《交通研究*C*辑：新兴技术》，19(6)：1306–1318。Fr¨uhwirth-Schnatter, S.（1994年）。数据增广与动态线性模型。《时间序列分析杂志》，15(2)：183–202。Gamerman, D. 和 Migon, H. S.（1993年）。动态层次模型。《皇家统计学会杂志：*B*辑（方法论）》，55(3)：629–642。Gilks, W. R.、Richardson, S. 和 Spiegelhalter, D.（1995年）。《马尔可夫链蒙特卡罗实践》。CRC出版社。Harvey, A. C.（1990年）。预测、结构时间序列模型与卡尔曼滤波器。

Lai, Y.、Golightly, A. 和 Boys, R. J.（2020年）。温度与湿度数据时空模型的序贯贝叶斯推断。《计算科学杂志》，43：101125。Petris, G.、Petrone, S. 和 Campagnoli, P.（2009年）。动态线性模型。载于《*R*语言动态线性模型》，第31–84页。施普林格。Rasmussen, C. E. 和 Williams, C. K. I.（2005年）。《机器学习中的高斯过程》。麻省理工学院出版社。Roberts, G. O. 和 Rosenthal, J. S.（2001年）。各种 Metropolis-Hastings 算法的最优尺度调整。《统计科学》，16(4)：351–367。Shaddick, G. 和 Wakefield, J.（2002年）。多个地点的每日多变量污染物数据建模。《皇家统计学会杂志：*C*辑（应用统计）》，51(3)：351–372。West, M. 和 Harrison, J.（2006年）。《贝叶斯预测与动态模型》。施普林格科学与商业媒体。WHO（2020年）。道路交通伤害。https://www.who.int/news-room/fact-sheets/detail/ road-traffic-injuries。 [在线；访问日期：24-May-2021]。