



Integrating a non-gridded space representation into a graph neural networks model for citywide short-term crash risk prediction

将非网格空间表示集成到图神经网络模型中，用于全市短期碰撞风险预测

# 主要内容

Main Contents

1

摘要

2

背景

3

方法

4

实验

5

结论



# 摘要



- 本文应用门控局部扩散图神经网络 (GLDNet) 模型来比较网格块(MB) 和网格这两种替代地理单元的使用，以预测未来时间窗口可能发生碰撞的位置。
- 测试是在一年的时间里使用来自澳大利亚墨尔本市的碰撞数据执行的。结果表明GLDNet始终优于基线方法。在地理单位方面，基于 MB 的 GLDNet 的性能优于其网格对应物。关于其适用性，基于 MB 的 GLDNet 直接与其他数据源集成，提供有关碰撞热点的上下文信息，有助于决策者制定警察巡逻和救援策略。



# 背景



**要解决的问题：** 根据碰撞发生的历史数据预测未来时间窗口内可能发生碰撞的位置。

**当前方法的缺陷：**

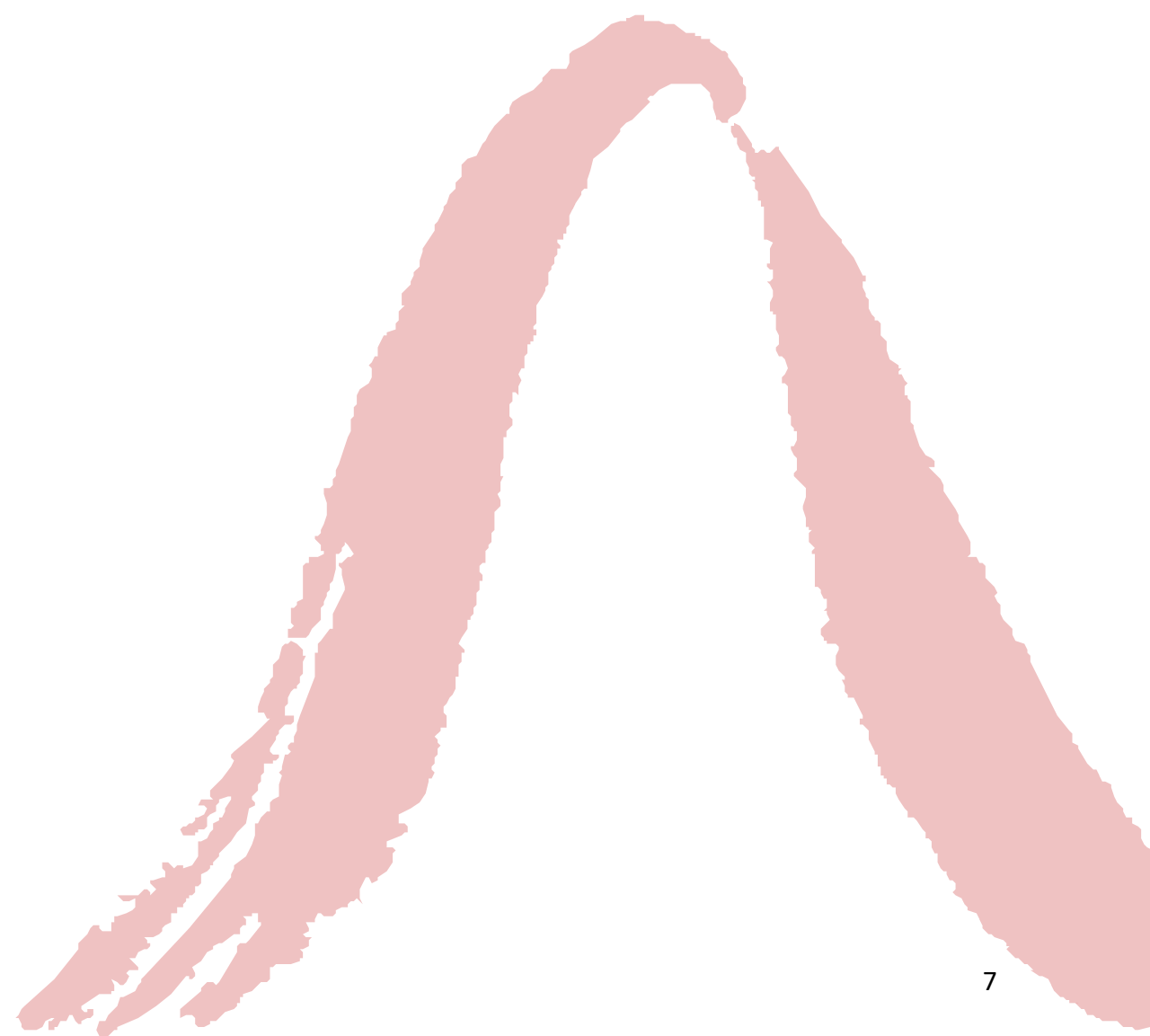
- 标准的深度学习方法受到将空间数据表示为网格的限制。标准的正方形网格不能恰好划分自然和建筑，会扭曲数据的空间相关性；研究区域边界的网格单元在形状或大小上通常与正常网格不同，可能会导致信息丢失和模型性能恶化。
- 当前考虑了多个数据源来预测全市的短期车祸风险。虽然辅助数据源提高了风险预测的质量，但它们也限制了GNN模型在现实世界中应用的可行性，因为它们增加了计算能力要求，并要求高精细时间分辨率的数据，难以获取。

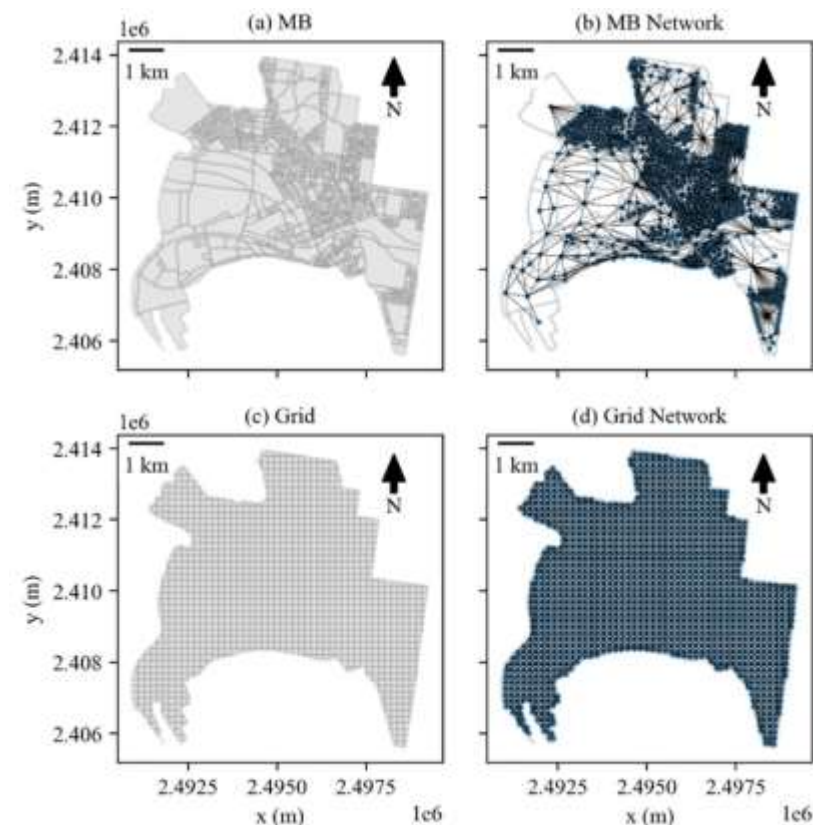
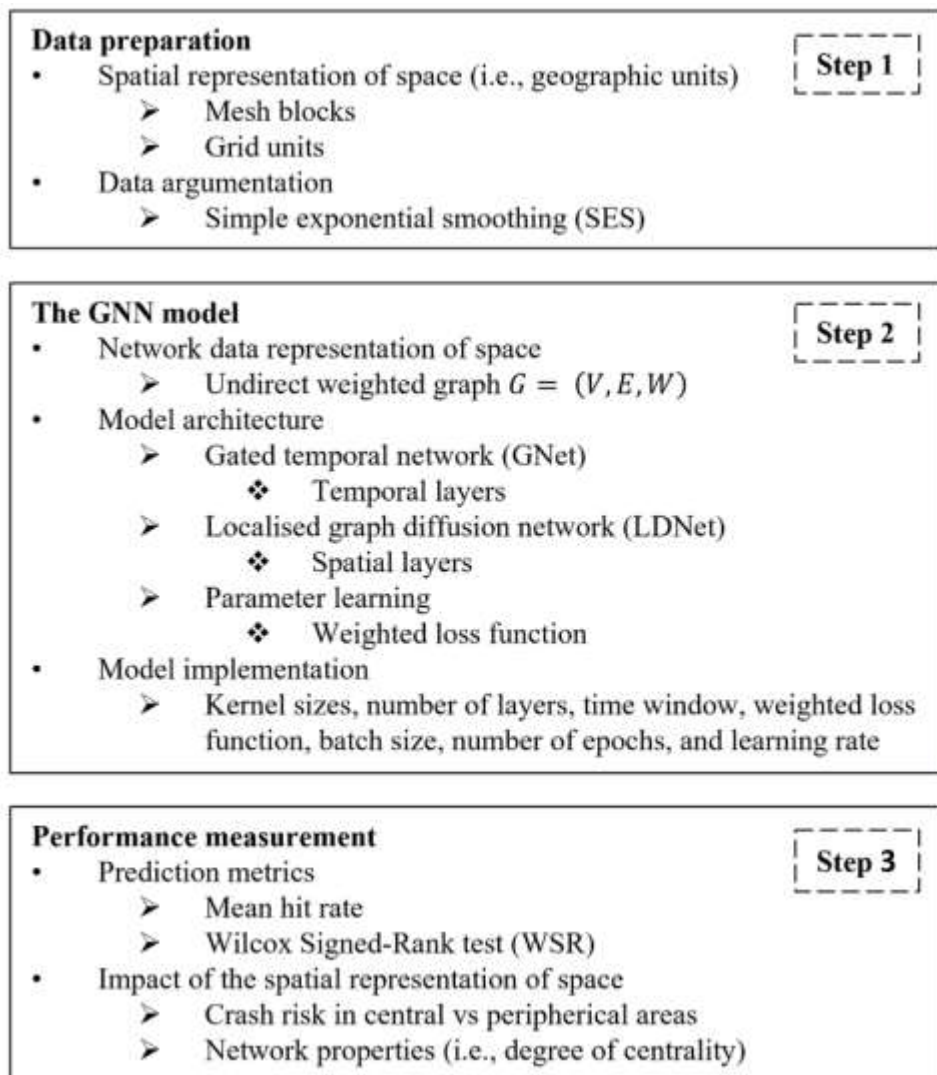
**本文：**

- 使用图神经网络（GNN）对图进行操作，不受空间网格表示的限制，它们依赖于基于网络的数据表示。对墨尔本在网格块级别(MB)和网格级别的城市进行了短期碰撞风险的预测。选择MB单元是因为它被决策者广泛使用来分配大量的城市资源，它是由澳大利亚统计局定义的最小地理区域（ABS）。
- 唯一的数据输入是历史碰撞发生信息。



# 方法





(a) 和 (c) 显示了研究区域的空间单元，而图 (b) 和 (d) 显示了它们相关的网络数据表示

分析框架



## ➤ 数据增强

使用简单的指数平滑（SES）技术来增强时域中的数据： $s_t = a_s x_t + (1 - a_s) s_{t-1}$ .

其中  $a_s$  是平滑因子，定义在0到1之间。更小的  $a_s$  导致更平滑的增广数据。在本文中，我们设  $a_s$  等于0.5。

## ➤ 网络数据表示

$G=(V,E,W)$  V、E和W分别是图节点、边和权重矩阵

无向图G用于表示数据并预测未来时间步长内碰撞发生的概率分布。无向图将地理单元表示为一组图节点，而边表示两个地理单元i和j是否相邻。权重矩阵  $w_{ij} \in W$  表示两个相邻地理单元  $e_{ij}$  之间的关系，并根据对象之间的相似性随着空间距离的增加而衰减的想法来定义。根据高斯核函数，边缘权重  $w_{ij}$  被定义为与两个相邻区域质心之间的欧几里得距离成反比：

$$w_{ij} = \begin{cases} \frac{\exp\left(-\left(\frac{\text{dist}_{i,j}}{h_i}\right)^2\right)}{\sqrt{2\pi}}, & \text{if } e_{ij} \text{ is an edge} \\ 0 & \text{otherwise,} \end{cases}$$

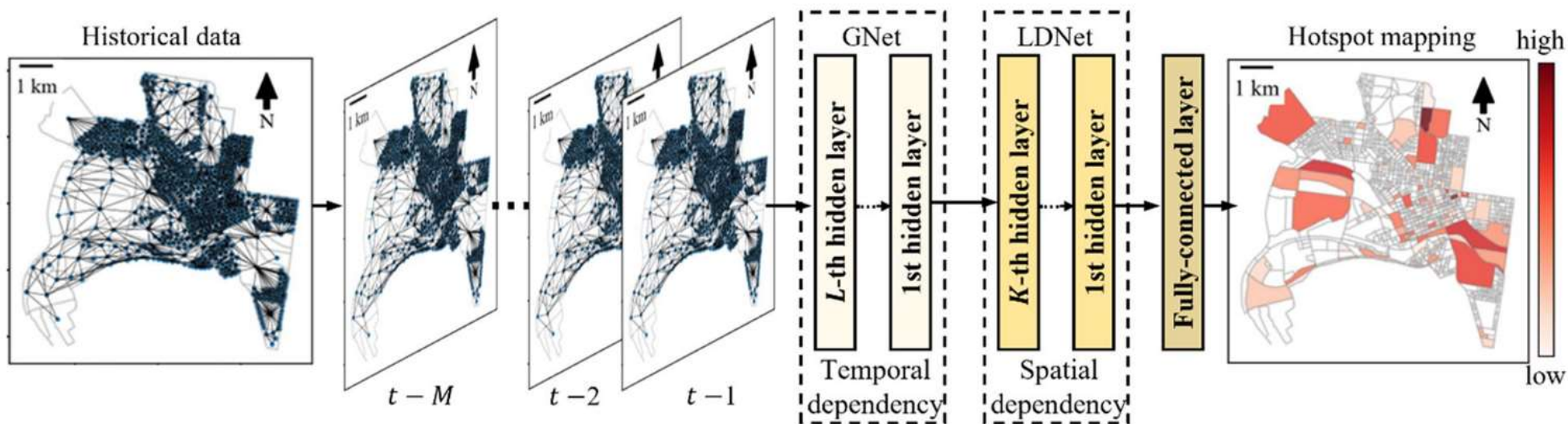
其中  $h_i$  是内核带宽， $\text{dist}_{ij}$  是两个相邻地理单元 i 和 j 的质心之间的距离。

## ➤ 模型架构

模型GLDNet是通过集成局部图扩散网络(LDNet)层、门控时间卷积网络(GNet)层和全连接层来定义的。

GNet 隐藏层学习历史崩溃事件随时间传播的影响。LDNet 隐藏层学习崩溃如何在空间中传播。全连接层将其转换为预测映射。映射  $GLDNet(X) = g_K W_{fc} + b_{fc}$  的碰撞发生概率。崩溃模型表述为：

其中  $g_K \in \mathbb{R}^{N \times m^k}$  是扩散网络的第k个隐藏层， $W_{fc} \in \mathbb{R}^{m^k}$  和  $b_{fc} \in \mathbb{R}^N$  是全连接层可学习参数。



GLDNet框架改编自Zhang和Cheng(2020)

## ➤ GNet组件

该组件是一个门控时间卷积网络 (TCN)，用于模拟碰撞事件的时间传播。该网络利用 ReLU 风格的门控机制，只有一个门，由 sigmoid 激活函数定义，由  $L$  个隐藏层组成。第  $l$  个表示为：

$$X^{l+1} = h_l(X^l) = \text{ReLU}(X^l W^l + b^l) \odot \sigma(X^l V^l + c^l) + d^l$$

where  $W^l, V^l \in \mathbb{R}^{n^{l-1} \times n^l}$ ,  $b^l, c^l \in \mathbb{R}^{n^l}$  and  $d^l \in \mathbb{R}^{n^l}$  are learnable parameters,  $\text{ReLU} = \max(0, x)$ , and  $\sigma(x) = 1 / (1 + \exp(-x))$  is the sigmoid function. The output  $X^{l+1}$  of

门控网络通过多层叠加定义为：

$$GNet(X) = h_L(h_{L-1}(\cdots h_2(h_1(X)) \cdots))$$

## ➤ LDNet组件

崩溃跨空间的传播可以被认为是遵循扩散过程，在图G中有限的随机游走序列后达到平稳分布。第k个隐藏层被公式化为：

$$X^{k+1} = g_k(X^k) = \text{ReLU}(X^k * \theta^k + PX^k * \eta^k)$$

where  $\text{ReLU} = \max(0, x)$  is the activation function,  $X^k * \theta^k$  captures the dependency of each node itself and  $PX^k$  represents the one-step random walk.  $\theta^k$  and  $\eta^k \in \mathbb{R}^{N \times m^{k-1} \times m^k}$  are learnable parameters in  $k$ -th hidden layer. The localised graph diffusion network is

对于输入X，扩散网络定义为：

$$\text{LDNet}(X) = g_K(g_{K-1}(\cdots g_2(g(X)) \cdots))$$

### ➤ 参数学习

加权损失函数定义为:  $loss = \frac{1}{N} \sum_{i=0}^N \omega_i (\hat{y}_i - y_i)^2$

其中N是节点数,  $\hat{y}_i$ ,  $y_i$ 分别是第i个节点的预测值和观测值,  $\omega_i$ 是分配给每个平方误差的权重。

### ➤ 基线方法

GLDNet模型与常用于预测时空事件的四种基线方法进行比较, 基线方法包括时空图卷积网络 (STGCN)、历史平均值 (HA)、自回归综合移动平均值 (ARIMA) 和梯度增强回归树 (GBRT)。

STGCN: <https://blog.csdn.net/yilulvxing/article/details/109545468>

HA:

[https://blog.csdn.net/weixin\\_41194129/article/details/143085876?fromshare=blogdetail&sharetype=blogdetail&sharerId=143085876&sharerefer=PC&sharesource=weixin\\_52377307&sharefrom=from\\_link](https://blog.csdn.net/weixin_41194129/article/details/143085876?fromshare=blogdetail&sharetype=blogdetail&sharerId=143085876&sharerefer=PC&sharesource=weixin_52377307&sharefrom=from_link)

ARIMA: <https://blog.csdn.net/fengdu78/article/details/121347188>

GBRT: [https://blog.csdn.net/March\\_A/article/details/129032437](https://blog.csdn.net/March_A/article/details/129032437)

## ➤ 性能测量指标

命中率定义为热点位置准确捕获的事件数量除以事件总数，因此不受大量零的影响，适用于评估稀疏事件预测。在本文中，考虑了最大 30% 的地理单元) 来计算命中率。这是因为，在更高的覆盖率水平下，任何模型都倾向于具有高性能。命中率表示为：

$$HR = \frac{n_{si}}{N_i}$$

其中HR是命中率， $n_{si}$ 和 $N_i$ 分别是覆盖区域s内的崩溃数量和整个研究区域在一个时间窗口i内发生的崩溃总数。

尽管平均命中率可用于直接比较不同预测方法的结果，但结果的统计意义尚不清楚。因此，为了评估结果的显著性，使用Wilcoxon Signed Rank (WSR) 检验来评估用不同方法获得的预测是否存在统计学差异。WSR测试统计数据由下式给出：

$$W_{WSR} = \sum_{i=1}^N (\text{sgn}(y_{1,i} - y_{2,i}) \cdot R_i)$$

其中N是样本量，sgn是用于提取实数符号的符号函数， $y_{1,i}$ 和 $y_{2,i}$ 分别是模型1和2中测试时间间隔i的命中率。 $R_i$ 是差值 $y_{1,i} - y_{2,i}$ 的秩。W WSR的统计显著性是使用单尾查找表获得的。



## ➤ 空间表示对GLDNet性能的影响

① 对中心和外围地区进行了评估

② 根据空间信息（即土地利用、人口）和网络属性（即网络密度和中心度）进行了解释

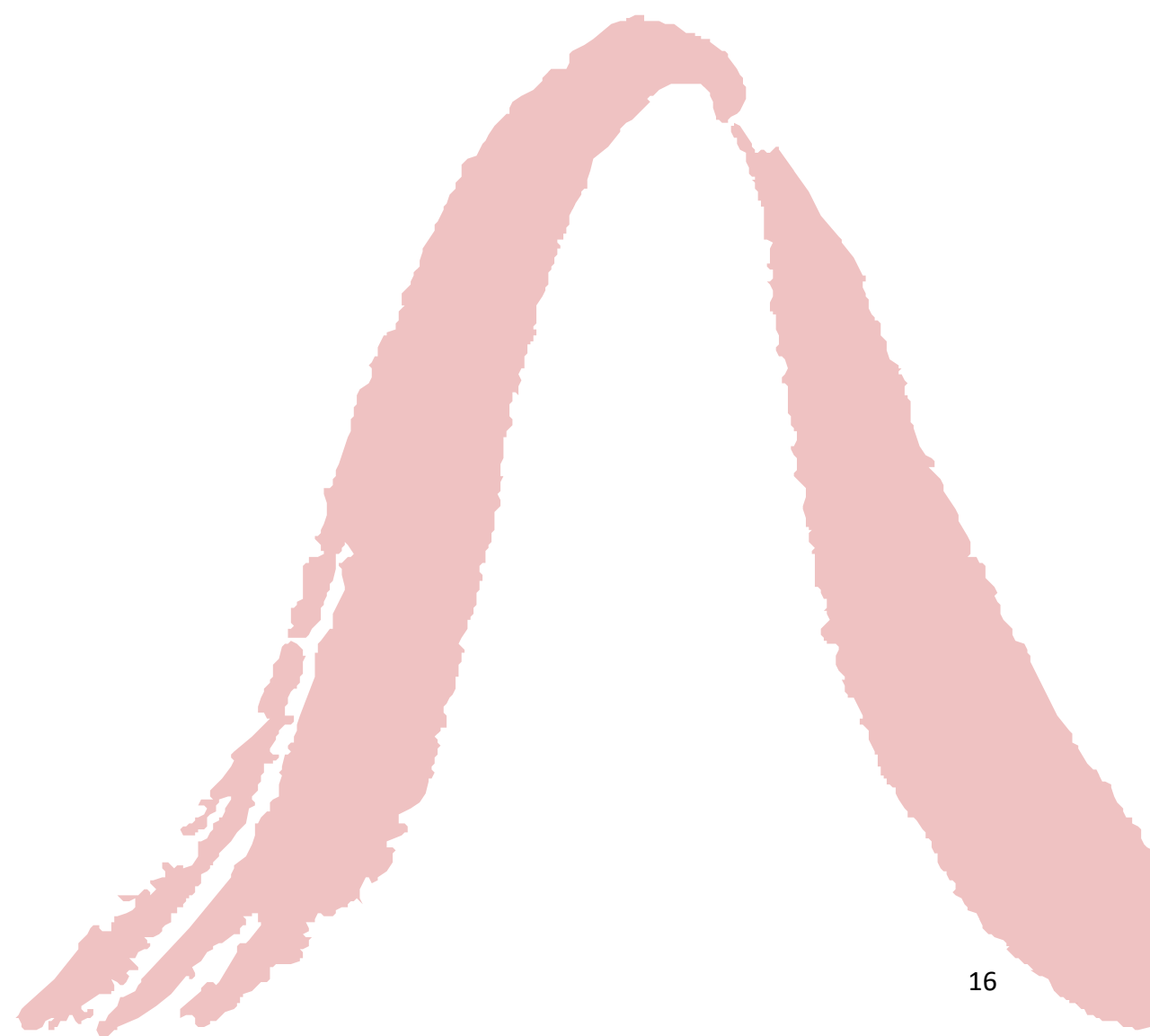
## ➤ 参数设置

设置高斯核函数来构建两个邻居地理单元之间相似性的权重矩阵，该高斯核函数的固定带宽 $h$ 等于所有地理单元最近邻居之间的最大距离， $MB$ 和网格分别为0.80km和0.17km。

训练、验证和测试样本分别设置为60%（876个间隔）、15%（219个间隔）和25%（365个间隔）。



# 实验





## ➤ GLDNet模型与基线方法的比较

Coverage level	Spatial unit	GLDNet	STGCN	HA	ARIMA	GBRT	GLDNet vs. STGCN	
							Difference <sup>a</sup>	p-value
5%	MB	36.6%	32.0%	31.2%	28.6%	28.3%	4.5%	0.0421
10%		52.8%	43.4%	44.7%	39.5%	39.8%	9.4%	0.0004
15%		59.6%	49.1%	47.9%	49.3%	49.3%	10.5%	0.0003
20%		66.8%	62.9%	49.7%	51.7%	51.7%	4.0%	0.0142
25%		69.8%	64.9%	53.4%	54.3%	54.3%	4.9%	0.0196
30%		71.3%	67.8%	57.4%	56.8%	56.8%	3.5%	0.0826
5%	Grid	26.1%	27.4%	20.2%	24.0%	25.1%	-1.2%	0.6815
10%		40.5%	38.4%	38.5%	39.5%	39.1%	2.1%	0.2314
15%		53.3%	51.1%	43.2%	52.0%	52.0%	2.1%	0.2228
20%		62.0%	57.1%	43.2%	53.5%	53.5%	4.9%	0.0569
25%		66.6%	61.0%	45.7%	55.4%	55.4%	5.6%	0.0264
30%		67.4%	61.6%	52.2%	57.4%	57.4%	5.8%	0.023

<sup>a</sup> Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

基线方法和两个地理单元的 GLDNet 模型的平均命中率

### ➤ 空间单元选择对GLDNet性能的影响

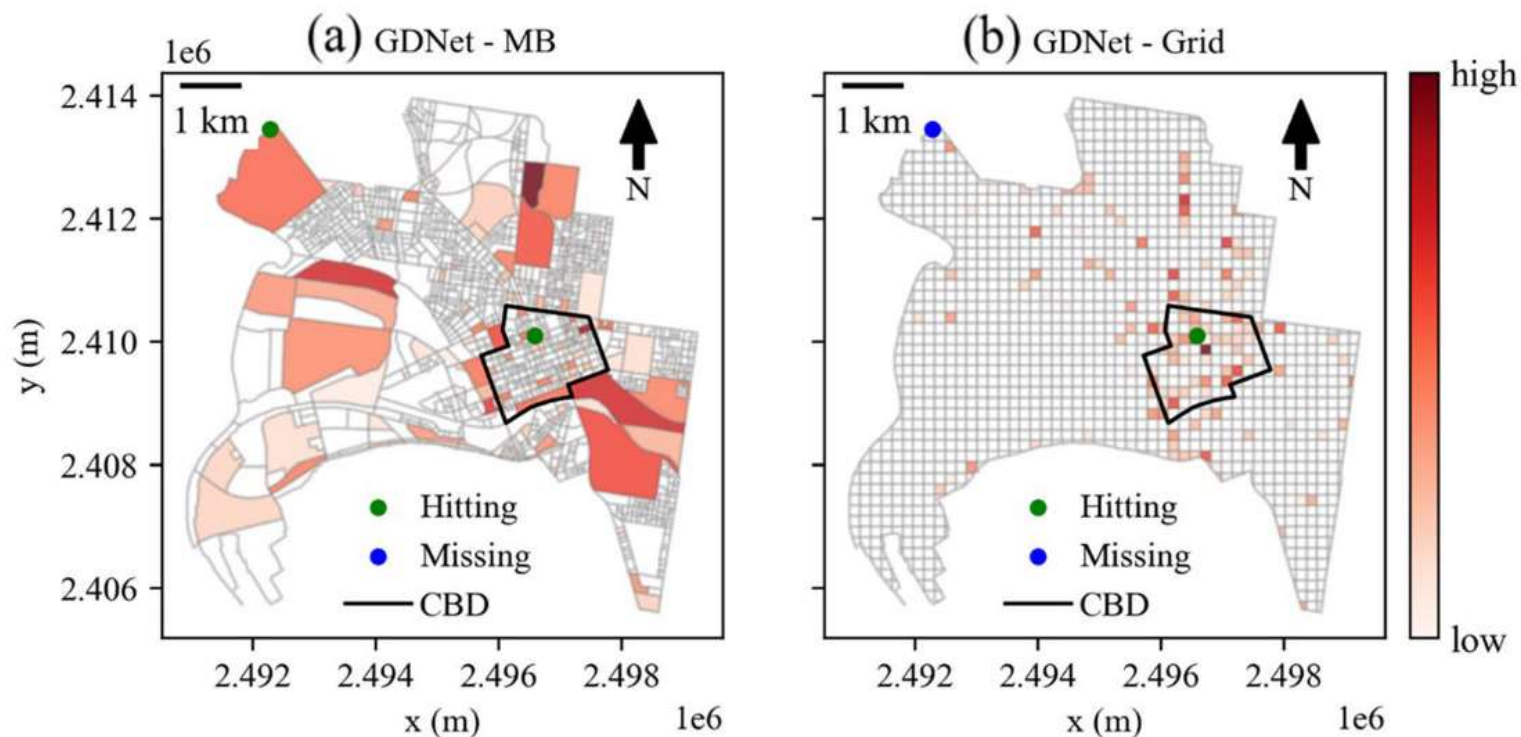
当覆盖率水平随着预测改进变得更加边际时，两种模型之间的差异都会下降，因为考虑了更多的区域。这解释了模型在 25% 和 30% 的覆盖率水平下表现在统计上不显著的差异。这些结果表明，与网格空间表示相比，采用行政地理单元来预测全市短期碰撞风险的潜在好处。

Coverage level	MB	Grid	Difference	<i>p</i> -value
5%	36.6%	26.1%	10.5%	0.0118
10%	52.8%	40.5%	12.3%	0.0046
15%	59.6%	53.3%	6.3%	0.0710
20%	66.8%	62.0%	4.9%	0.0803
25%	69.8%	66.6%	3.2%	0.1612
30%	71.3%	67.4%	3.9%	0.1139

基于两个地理单元的GLDNet模型的平均命中率比较

### ➤ 碰撞频率随空间变化的影响

基于MB的碰撞风险空间分布遍布墨尔本市，而基于网格的模型导致碰撞风险的空间分布集中在城市商业区（CBD）周围。



10% 覆盖水平下崩溃风险的预测映射（绿色和蓝色点分别代表捕获和未捕获的崩溃事件）

关于 **CBD 区域**，我们观察到基于网格的模型在所有覆盖级别上的平均命中率高于 MB 的结果。**非 CBD 区域**，我们观察到相反的情况。然而，CBD 和非 CBD 区域的基于 MB 和基于网格的模型之间的平均命中率差异很大。对于**整个研究区域**，基于 MB 的模型在所有覆盖级别上都优于其网格对应物，平均命中率更高。

Coverage level	CBD				Outside CBD			
	MB	Grid	Difference*	p-value	MB	Grid	Difference <sup>a</sup>	p-value
5%	11.6%	14.9%	-3.3%	0.0924	24.9%	11.2%	13.7%	0.0006
10%	13.5%	19.5%	-6.0%	0.0030	39.3%	21.0%	18.3%	0.0000
15%	15.8%	21.9%	-6.1%	0.0051	43.8%	31.3%	12.4%	0.0013
20%	18.4%	24.7%	-6.3%	0.0081	48.4%	37.2%	11.2%	0.0008
25%	19.4%	25.5%	-6.1%	0.0051	50.4%	41.1%	9.3%	0.0089
30%	19.8%	25.5%	-5.7%	0.0091	51.5%	41.9%	9.6%	0.0053

<sup>a</sup> Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

基于 CBD 和 CBD 区域外每个空间位置单元的 GLDNet 预测性能

基于MB的GLDNet捕获中心区域外碰撞风险的能力由两个主要因素解释：（1）与MB单元边界定义相关的基础信息和（2）网络属性。首先，土地利用、住宅数量和道路网络等空间信息被用于开发MB单元。这些变量也与碰撞发生的可能性相关，从而影响短期碰撞风险的预测。其次，关于网络属性，与网格表示相比，MB具有更高的平均中心度、紧密中心度、K核数和聚类系数。

MB也可能与高密度区域中空间单元的较低性能有关。每个MB单元的设计都是为了容纳类似数量的住宅。换句话说，MB空间表示的粒度很高，但碰撞发生的方差很低，这增加了预测误差。从这个意义上说，合并高密度地区的一些地理单元或不受住宅数量影响的替代地理单元的策略可能会提高GNN模型在碰撞热点测绘应用中的性能。

### ➤ 碰撞频率变化的影响

对于碰撞次数少于两次的间隔，基于MB的模型优于其网格模型。相反，基于网格的模型在更高的覆盖水平下，对于发生两次或两次以上碰撞的间隔，呈现出更高的平均命中率。

Coverage level	$0 \leq \text{crash count} < 2$				$2 \leq \text{crash count} < 6$			
	MB	Grid	Difference <sup>a</sup>	p-value	MB	Grid	Difference <sup>a</sup>	p-value
5%	38.9%	27.4%	11.6%	0.0315	31.1%	23.3%	7.8%	0.0797
10%	57.9%	42.1%	15.8%	0.0068	41.1%	36.8%	4.2%	0.2557
15%	65.3%	54.7%	10.5%	0.0385	46.5%	49.8%	-3.3%	0.5883
20%	71.6%	60.0%	11.6%	0.0205	55.9%	66.5%	-10.7%	0.8772
25%	74.7%	66.3%	8.4%	0.0721	58.4%	67.3%	-8.9%	0.8774
30%	75.8%	67.4%	8.4%	0.0653	60.8%	67.3%	-6.5%	0.8028

<sup>a</sup> Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

GLDNet模型在两个地理单元的每日碰撞计数水平下的平均命中率



### ➤ 模型适用性的影响

第一个热点位于市中心，包括中央商务区及其一些一级邻居。第二个热点主要是位于东南部两个郊区之间的大型公园。图b表明，大多数崩溃事件都集中在特定链路周围。第三个热点主要是公园。热点四和五主要是由大型MB定义的商业或工业区。热点4主要由几条主干道定义，而热点5由一条主要高速公路和几条主干路组成。第六个热点包括进入当地海滩和墨尔本港口。

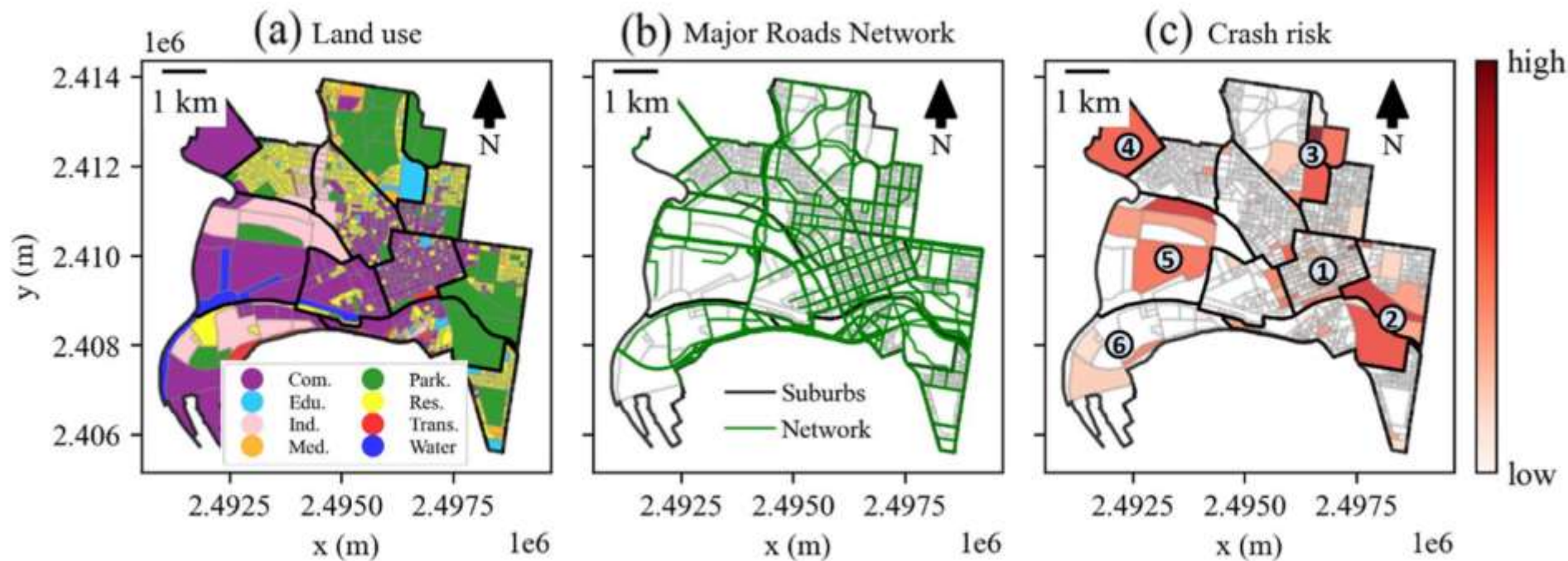



图 (c) : 10%覆盖水平下的碰撞风险预测图



# 结论



本研究采用了一个GNN模型GLDNet来预测全市的短期碰撞风险。与之前用于碰撞风险预测的GNN应用相比，所实现的模型将碰撞事件的历史发生视为唯一的数据输入，不受空间网格表示的约束，这大大提高了交通管理和警察执法机构从业者的适用性。

**谢谢**

