# 爱尔兰表土污染分析：一种聚类方法

Mimi Zhang

School of Computer Science and Statistics, Trinity College Dublin, Ireland



## 1 背景

2023 年 7 月，欧盟委员会提出了土壤监测法，以标准化整个欧盟的系统性土壤监测，并规定创建国家污染场地登记册 (European Commission, 2023)。然而，爱尔兰面临着在五年内实施该法规的重大挑战，主要是由于缺乏一个集中的国家级登记册——这是立法的核心要求。此外，不到三分之一的爱尔兰土壤研究与 2030 年欧盟土壤战略和国家政策中概述的研究重点相一致 (McNamara et al., 2022)。将爱尔兰研究与欧洲标准进行比较表明，需要采用更标准化的方法来评估土壤污染，以符合新的监管框架。

本研究分析了由爱尔兰地质调查局（GSI）领导的国家测绘计划 Tellus 项目 (Tellus, 2019) 浅层表土地球化学数据。Tellus 数据集为解决爱尔兰土壤污染知识空白提供了一个宝贵但未充分利用的资源。尽管之前的用途仅限于单元素分析或局部区域研究，本研究采用了多变量和机器学习技术来揭示潜在有毒元素（PTEs）之间的复杂相互作用。用于生成聚类的数据和可重现代码在 GitHub 存储库中公开获取：`https://github.com/tobinjo96/CPFcluster/tree/master/Spatial-CPF`。由于文件大小限制，无法直接托管将土壤样品可视化到基岩和土地覆盖地图上的数据集和脚本，但可根据请求提供。

## 2 数据准备

该数据集被称为 “G5”，是在 2017 年至 2019 年间收集的，并且公开可用这里。它覆盖了 17,983 平方公里（占国家总面积的 24.3%），涵盖了爱尔兰的西部、中部和东部地区；参见图 1。这些地区的地质构造和土地使用多样，对 PTEs 的分布和移动有显著影响。值得注意的是，之前的研究主要集中在都柏林和北爱尔兰 (McNamara et al., 2022)，而其他地区则研究不足。本研究通过利用 GSI Tellus 数据集来弥补这一空白。该数据集包括在每 4 平方公里一个采样点、深度为 5-20 厘米处采集的 4,278 个表土样本。样品经过了多元素部分

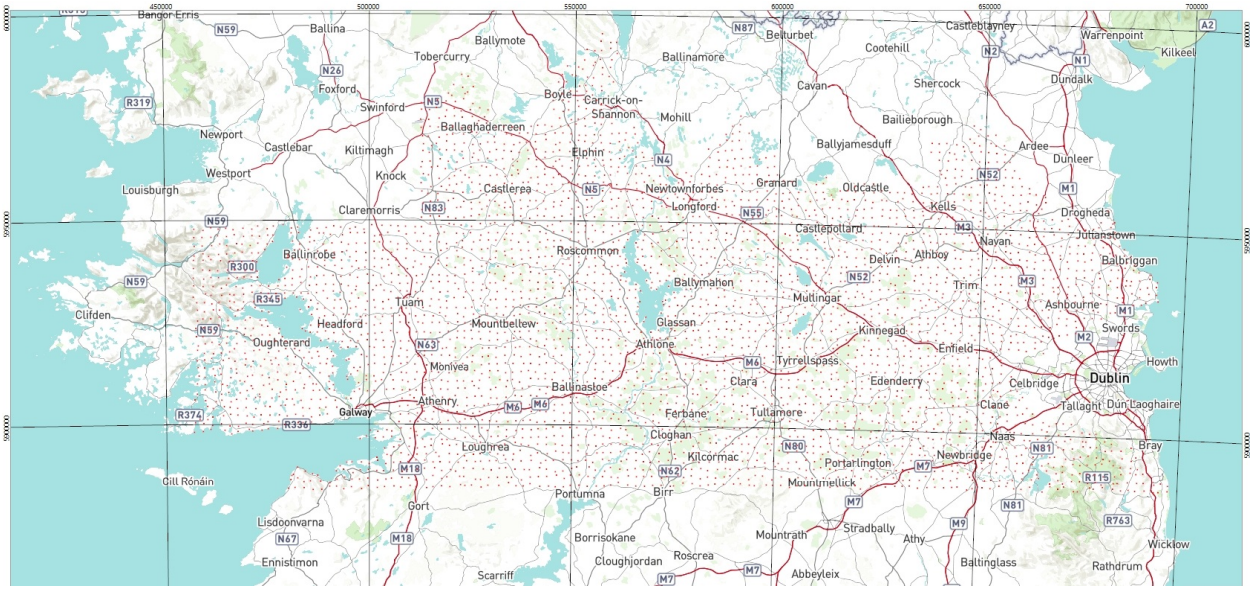


图 1: “G5” 数据集的采样点（红点）。

提取分析，使用了电感耦合等离子体质谱（ICP-MS）和王水消化处理，在认证设施中进行。在 ICP-MS 分析过程中，由于某些元素（Ta、Au、Pd、Pt 和 Re）的值低于检测限 $> 5\%$，因此这些元素被排除在外。对于本研究而言，选择了 15 种关键 PTEs（以 mg/kg 为单位测量）：As、Ba、Bi、Co、Cr、Cu、Mn、Mo、Ni、Pb、Sb、Sn、U、V 和 Zn。这些元素的选择基于三个标准：（1）其在土壤中较高的污染风险 (Reimann et al., 2018)，（2）对人类健康和环境的已记录有害影响，以及（3）在爱尔兰土壤中的历史普遍存在性。这种有针对性的选择确保了对威胁土壤质量的最大危害因素进行集中分析。

坐标从爱尔兰横墨卡托(ITM,EPSG:2157)系统转换为全球 WGS84 系统(EPSG:4326)，以确保与网络地图库（例如 Folium）和 GIS 工具的兼容性。原始数据集和处理后的数据集均可在 `https://github.com/tobinjo96/CPFcluster/tree/master/Spatial-CPF/Data` 获得。从 GSI 网站获取了一份比例尺为 1:1,000,000 的基岩地质图（Shapefile 格式）。土地覆盖地图是使用欧洲联盟哥白尼陆地监测服务的数据生成的，参考年份为 2018。

# 3 方法论

在这项研究中，我们应用 CPF 聚类方法 (Tobin and Zhang, 2024) 将土壤样本分类为不同的组，并分析爱尔兰表土污染模式。该方法在公开的 Python 包 CPFcluster（`https://github.com/tobinjo96/CPFcluster`）中实现。鉴于我们的数据集中包含了地理坐标，我们使用 Spatial-CPF 文件夹中的函数，在聚类过程中纳入了空间约束。

我们首先运行创建邻接矩阵函数，从地理坐标构建邻接矩阵（即互为 k 近邻图）。由于其体积较大，预计算的邻接矩阵文件 *geo_邻域_邻接矩阵.npy* 未包含在 Data 文件夹中。用户必须运行提供的代码来生成该矩阵。返回的邻接矩阵和地球化学污染数据随后被输入到 *cpf.fit* 函数中。[1] 根据 Tobin and Zhang (2024) 的建议，我们在 $min\_samples = \sqrt{n} \approx 70$ 附近调查超参数空间。通过广泛的参数配置测试，我们确定了最优的超参数集：*min_samples=75, rho=0.01, alpha=0.015, merge_threshold=7.5, density_ratio_threshold=0.7*，在此情况下 Calinski-Harabasz 分数为 76.5202。用户可以通过运行代码_示例_几何.*py* 脚本来测试替代的超参数。聚类结果使用降维技术在图 2 中进行了可视化：

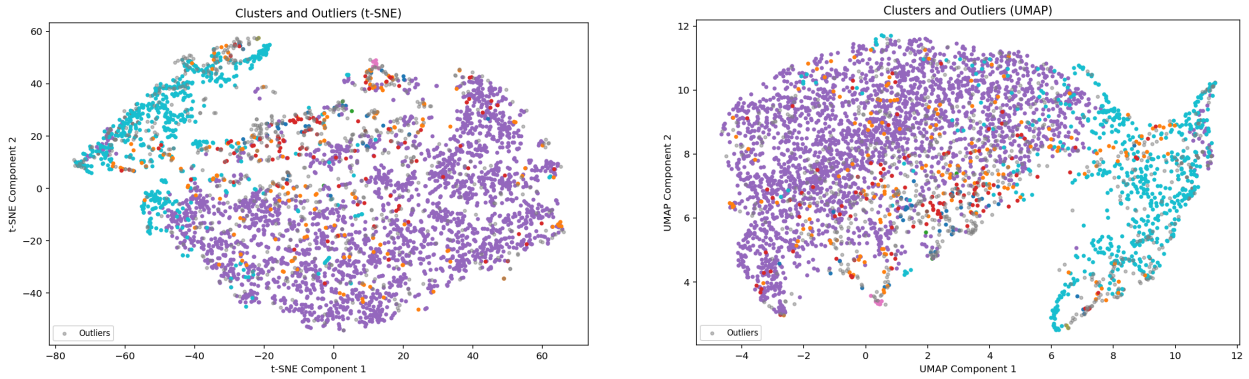


图 2: 通过 t-SNE 投影（左）和 UMAP 投影（右）可视化聚类结果。不同颜色代表不同的聚类。

# 4 结果

CPF 算法产生了八个聚类和一个包含 682 个离群点的集合，聚类大小详情见表 1。聚类 1（用紫色点表示在图 2 中）构成主要群体并含有健康的土壤样本，而较小的聚类（7 和 8）以及离群点集因它们潜在揭示异常土壤条件的能力而特别值得关注，这将在以下部分进行分析。

图 3 比较了三个最大子集中的 PTE 浓度：聚类 1、聚类 2 和异常值集合，y 轴采用对数转换以适应异常值集合中极端的 Mn 和 Zn 值。如图 3 所示，聚类 1 表现出最小的四分位间距（IQR），表明在三个子集中 PTE 浓度的变化性最低。聚类 2 显示出最低的中位数和最低的上须，表明浓度一直较低。相比之下，异常值集合显示了最高的中位数和最大的 IQR，反映出高浓度且变化性大，并伴有众多极端的上须。

[1] 构建_*CC* 图在核心_地理文件中将仅基于地球化学污染数据构建另一个邻接矩阵，应用与空间邻近矩阵相同的 k-最近邻阈值。最终的邻接矩阵是这两个矩阵（地理和地球化学）的按元素乘积计算得出。因此，只有当两个样本在地球化学上相似且地理位置接近时，才被认为是邻居。最终的邻接矩阵用于识别连通分量。

表 1: G5 污染数据的聚类结果

| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Size | 2623 | 604 | 173 | 114 |
| | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 |
| Size | 40 | 25 | 9 | 8 |

图 4 和 5 比较了八个土壤群组中每个元素的箱线图。关键观察结果包括：

- 集群 2 和集群 6 以元素 Mn、Ba、Zn 和 Pb 为特征，这些元素的浓度模式与其他集群明显不同。虽然 Mn、Ba 和 Zn 在所有其他集群中表现出高浓度，但在集群 2 和集群 6 中的值相对较低。相比之下，Pb 在整个集群（包括异常集）中保持一致的浓度。此外，在集群 2 和集群 6 中，这四种元素（Mn、Ba、Zn 和 Pb）的浓度显著高于剩余的 11 种元素。图 6 显示了集群 2 和集群 6 土壤样本的采样地点地理分布。图 6 表明低浓度的土壤样本主要集中在三个地区。如土地覆盖地图所示，集群 2 和集群 6 中的土壤样本来自湿地和农业区。中部区域的新月形特征需要进一步调查。

- 集群 3、4 和 5 显示出相似的元素模式，其中集群 5 通过砷 (As)、钡 (Ba)、钼 (Mo)、锑（Sb）和铀（U）的较高中位浓度和更大的变异性而脱颖而出。图 7 显示了三个集群中土壤样本的地理分布。集群 3 主要位于奥法利郡，有少量延伸至戈尔韦郡，而集群 4 则集中在威克洛郡。集群 5 呈现出明显的月牙形空间特征。

- 集群 7 以铜 (Cu)、铅 (Pb) 和锡 (Sn) 为特征，它们的中位浓度显著更高且变异性更大，与在其他集群中的模式相比。集群 8 中的砷 (As) 和铀 (U) 浓度表现出与其他集群相比显著的变异性。集群 7 和 8 的土壤样品如图 8 所示，其中第二张和第三张地图分别是基岩地图和土地覆盖地图。集群 7 和 8 共包含 17 个样本。这些样本中大多数（15 个）是基岩石灰石，其余两个被识别为深海灰岩，一种砂岩类型。这些土壤样品来自以湿地和农业用地为主的区域。

分析根据 15 种 PTEs 的空间分布模式将其分为两个不同的组。第一组包括 Ba、Cu、Mn、Zn、As、Mo、Pb、Sb、Sn 和 U，在调查区域的不同地区显示出浓度的显著变化。相比之下，第二组包括 Cr、Ni、V、Bi 和 Co，在整个研究区域内保持一致的浓度水平。

并非离群值集合中的所有样本都具有极高的浓度；有些被标记为离群值是因为它们的多元素组成与相邻样本有显著差异。为了仅基于浓度（独立于地理因素）识别异常样本，我们将孤立森林方法应用于离群值集合，并将污染参数设置为 30%。该方法识别出 205 个离群值，所有这些都表现出一种或多种 PTE 的极高水平。在图 9 中，上图显示了离群值集合中的所有土壤样本，而中图和下图则专注于孤立森林检测到的 205 个异常值。这些离群值的空间分布与整个离群值集合非常相似，没有嵌套模式表明可能存在大规模污染。然而，

中图和下图揭示了在经度 9° 至 9.5°（爱尔兰以西）和 7.5° 至 8° 之间的区域显示出相对较高的污染水平。

# 5 总结

本研究利用 Tellus 计划的地球化学数据，分析了爱尔兰境内 17,983 平方公里范围内的 4,278 个土壤样本，调查表层土壤污染情况。该研究采用空间-CPF 算法（具有空间约束条件的 CPF 聚类方法），将样本分为八个集群和一个包含 682 个土壤样本的异常值集，揭示了不同的污染模式。

关键发现包括：

- 聚类 1 代表毒性元素变异程度较低的健康土壤。
- 聚类 2 和 6 也代表健康的土壤。但它们表现出独特的 Mn、Ba、Zn 和 Pb 模式，与其它聚类不同。
- 集群 7 和 8（17 个样本）显示出铜、铅、锡、砷和铀的含量升高，这与石灰岩基岩和湿地/农业区域有关。
- 孤立森林识别出 205 个具有极高 PTE 浓度的极端离群值。地理热点出现在经度 9° 至 9.5°（爱尔兰西侧）和 7.5° 至 8° 之间，尽管没有检测到大规模污染模式。

# 致谢

我谨向 Áine Sadlier 表达我的感激之情，感谢她作为毕业项目的一部分所做的宝贵初步探索工作。

# References

European Commission (2023). Directive of the European Parliament and of the Council. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A52023PC0416 [Accessed: April/2025].

McNamara, M., Binner, H., Hynes, E., and Andrade, L. (2022). A Signpost for Soil Policy in Ireland. https://www.epa.ie/publications/research/evidence-synthesis-reports/Evidence_Synthesis_Report_1.pdf [Accessed: April/2025].

Reimann, C., Fabian, K., Birke, M., Filzmoser, P., Demetriades, A., Négrel, P., Oorts, K., Matschullat, J., and de Caritat, P. (2018). GEMAS: Establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. *Applied Geochemistry*, 88:302–318.

Tellus (2019). `https://www.gsi.ie/en-ie/data-and-maps/Pages/Geochemistry.aspx` [Accessed: April/2025].

Tobin, J. and Zhang, M. (2024). A theoretical analysis of density peaks clustering and the component-wise peak-finding algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1109–1120.

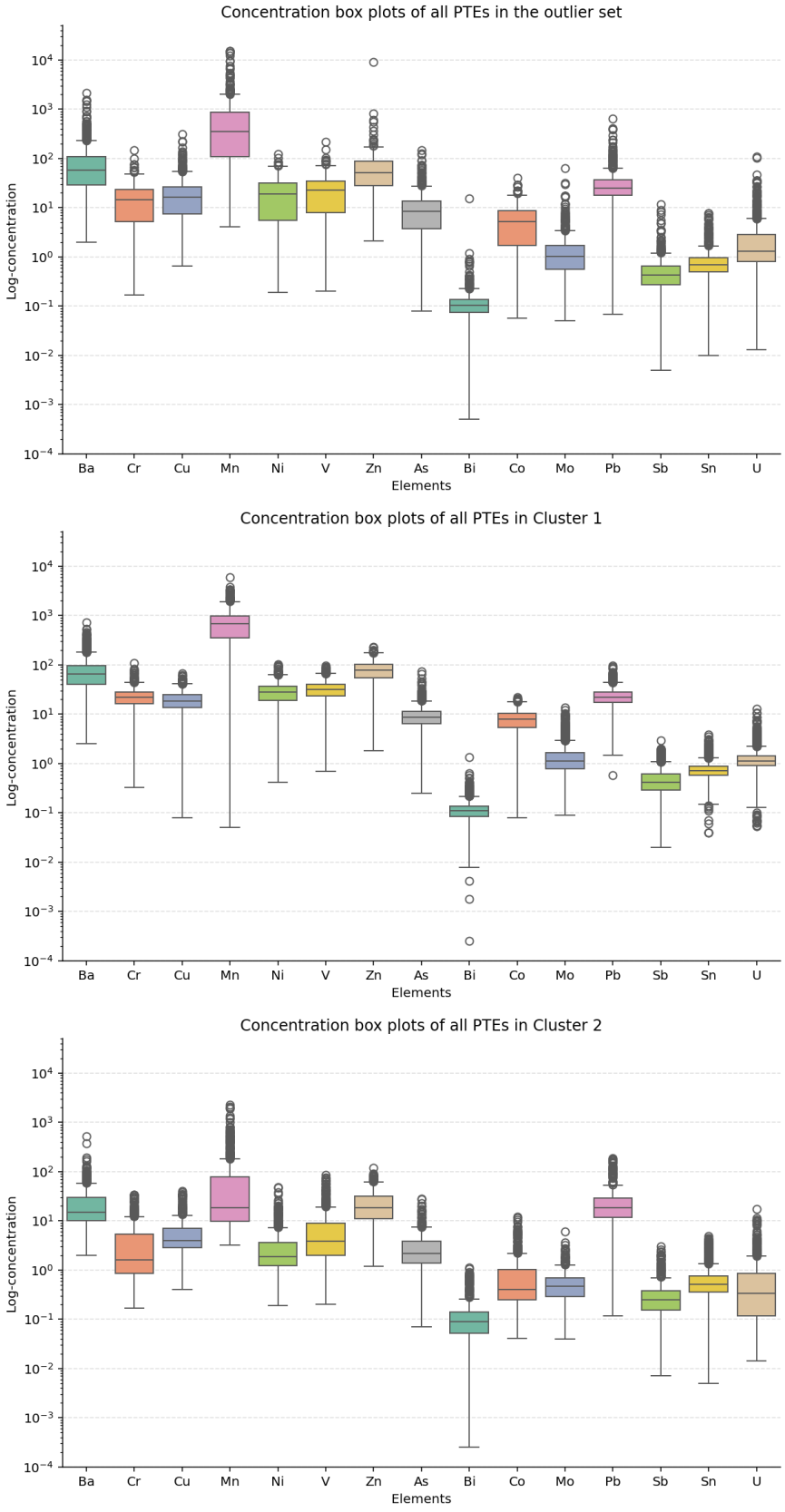


图 3: 异常值集（顶部）与聚类 1（底部）之间 PTE 浓度的比较。请注意对数 $_{10}$ 转换后的 y 轴比例，这是为了可视化异常值中的极端 Mn 浓度所必需的。

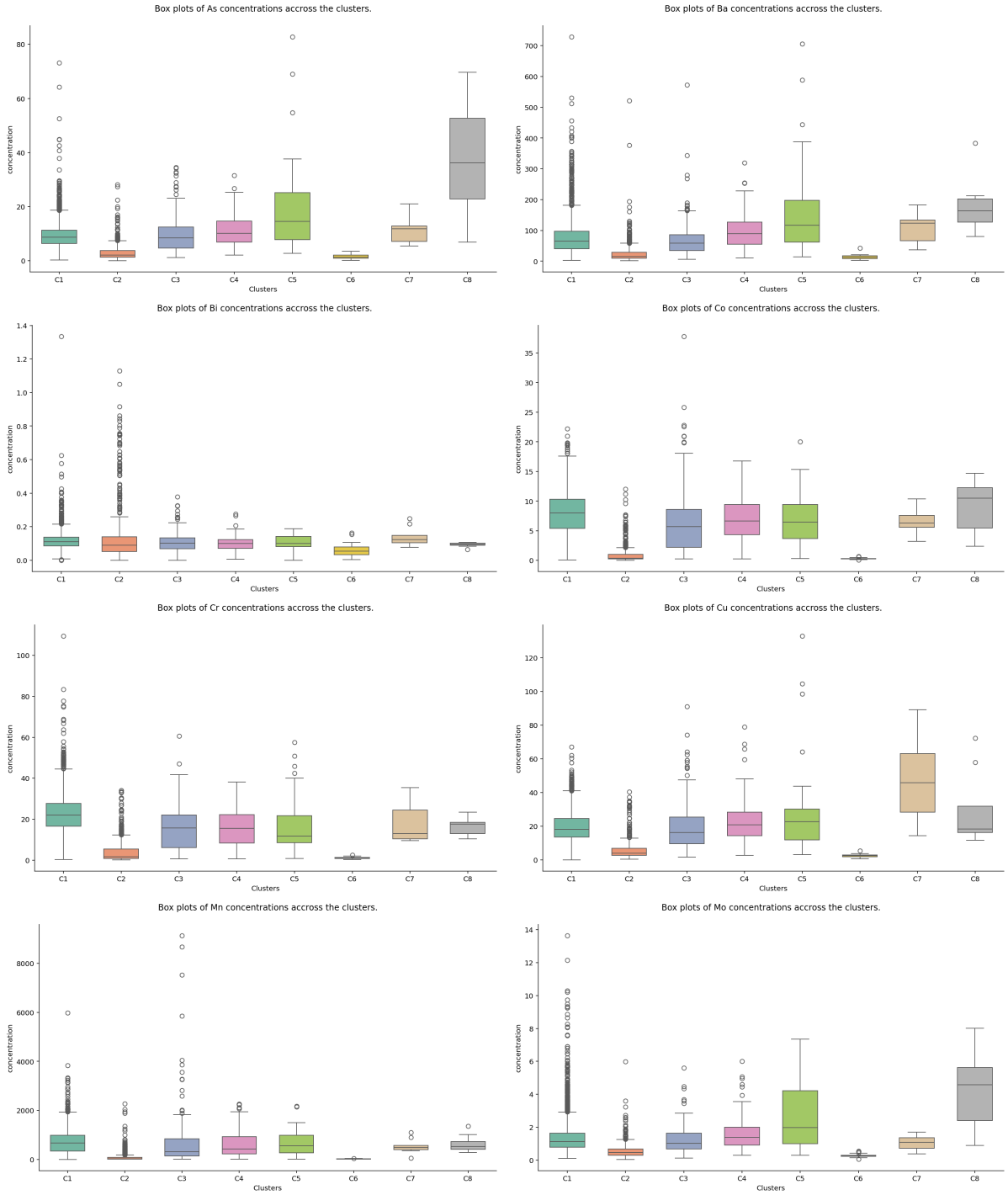


图 4: 显示八个识别土壤群集中 15 种 PTE 浓度分布的箱线图。

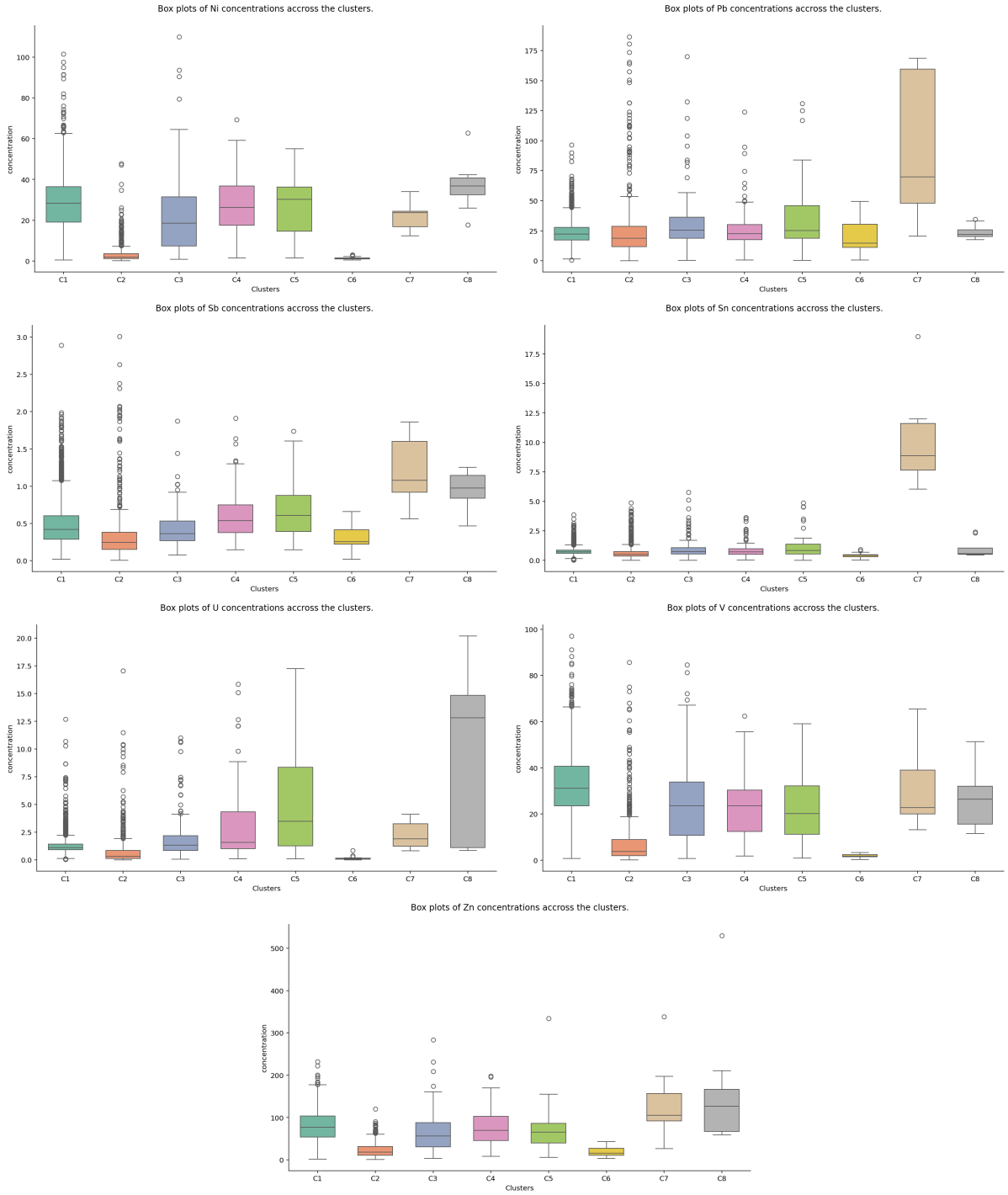


图 5: 箱线图显示了八类识别土壤中 15 种 PTE 浓度的分布。

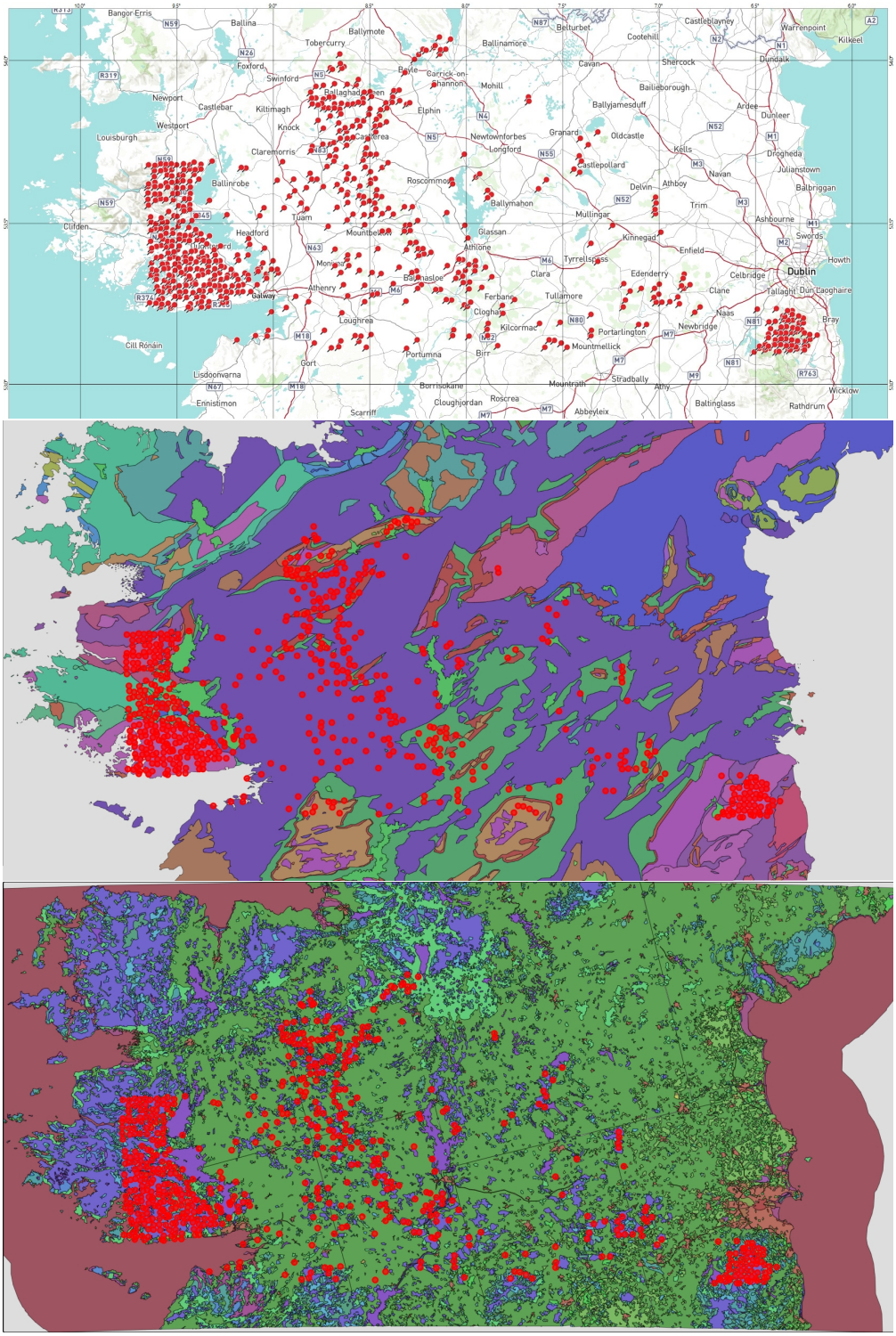


图 6: 集群 2 和集群 6 土壤样本的采样点地理分布。背景基岩类型主要是石灰岩、花岗岩和砂岩。背景土地覆盖类型主要是湿地和农业区。

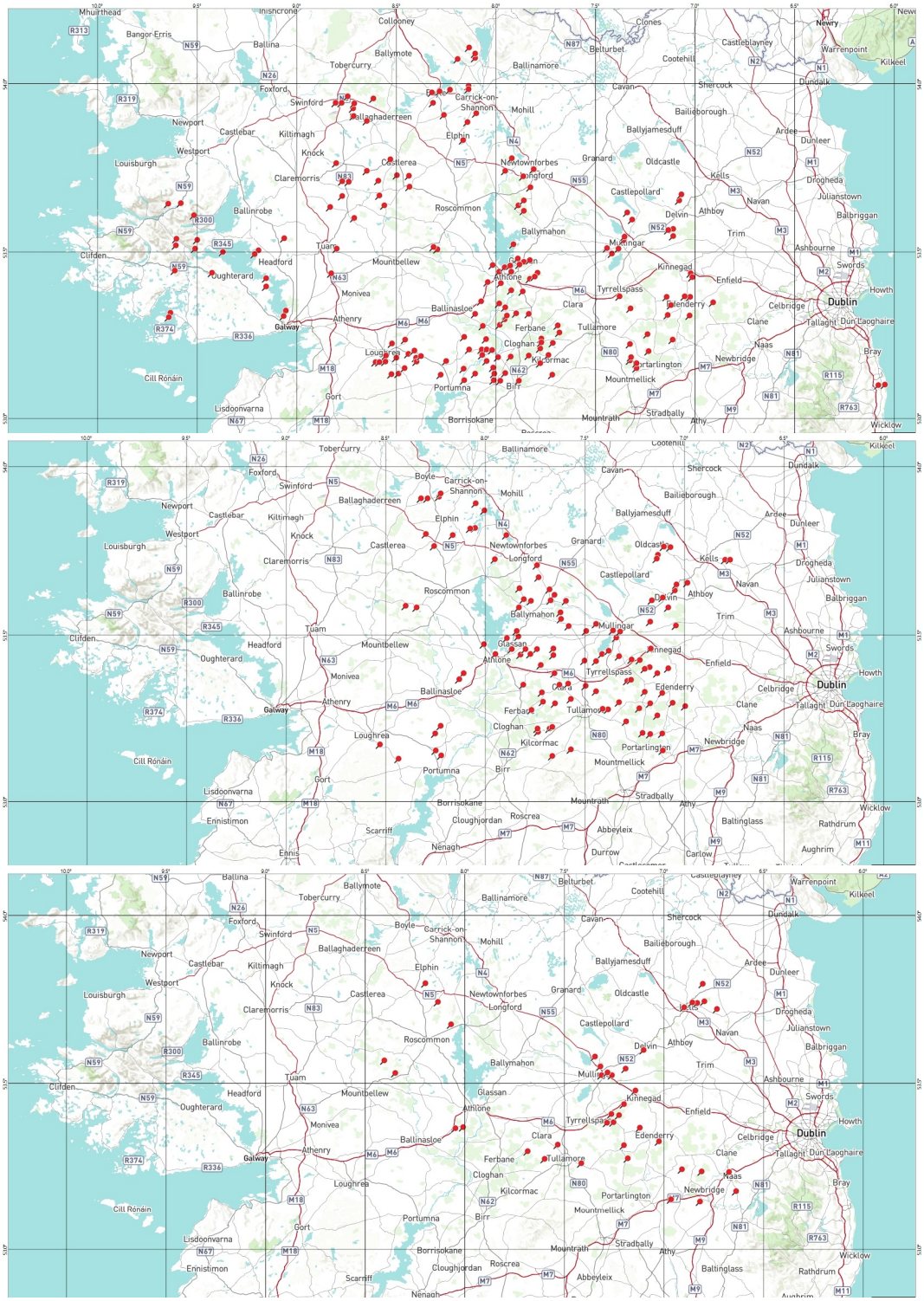

图 7: 集群 3（顶部）、集群 4（中部）和集群 5（底部）的土壤采样点地理分布。

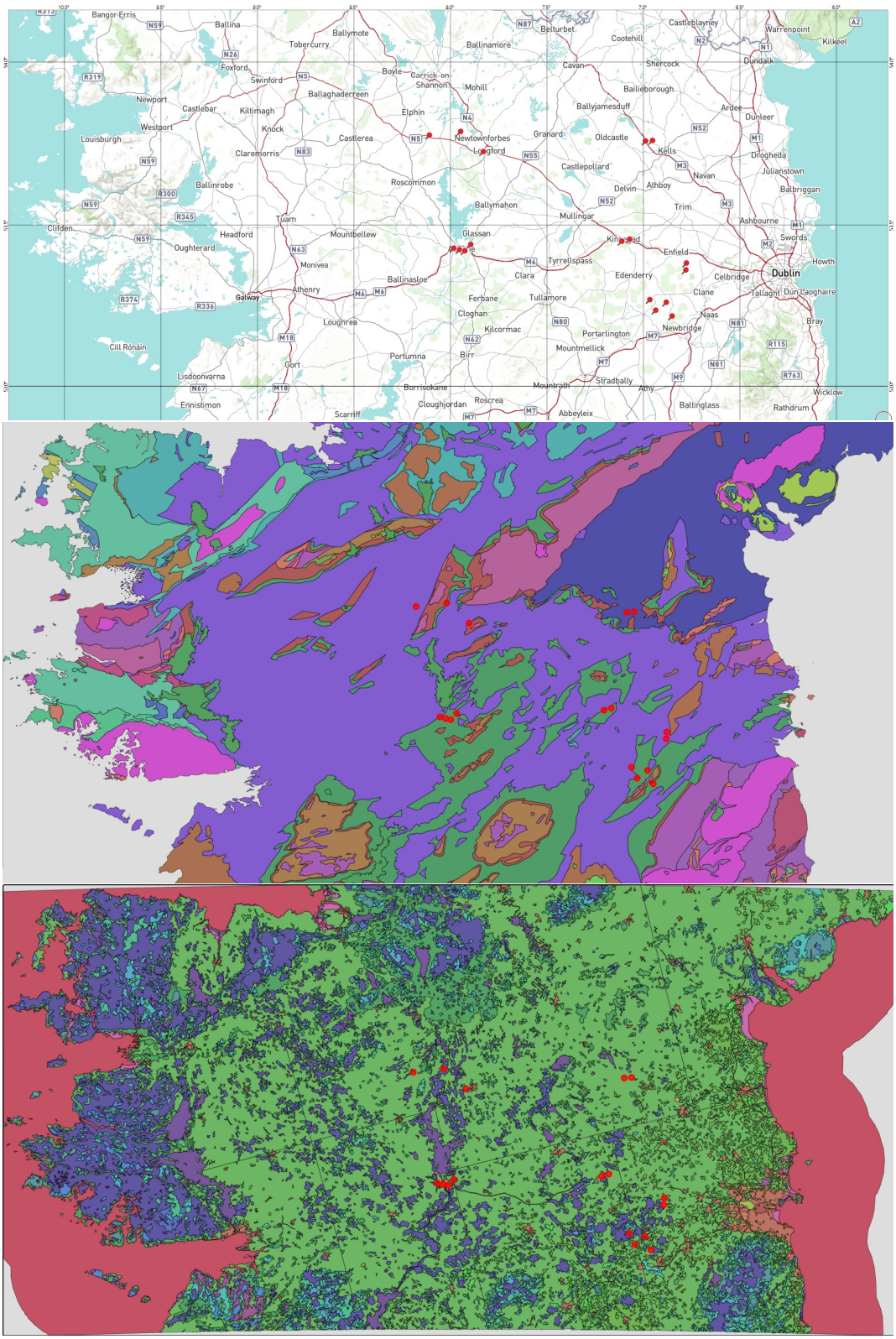


图 8: 集群 7 和集群 8 的采样点地理分布。

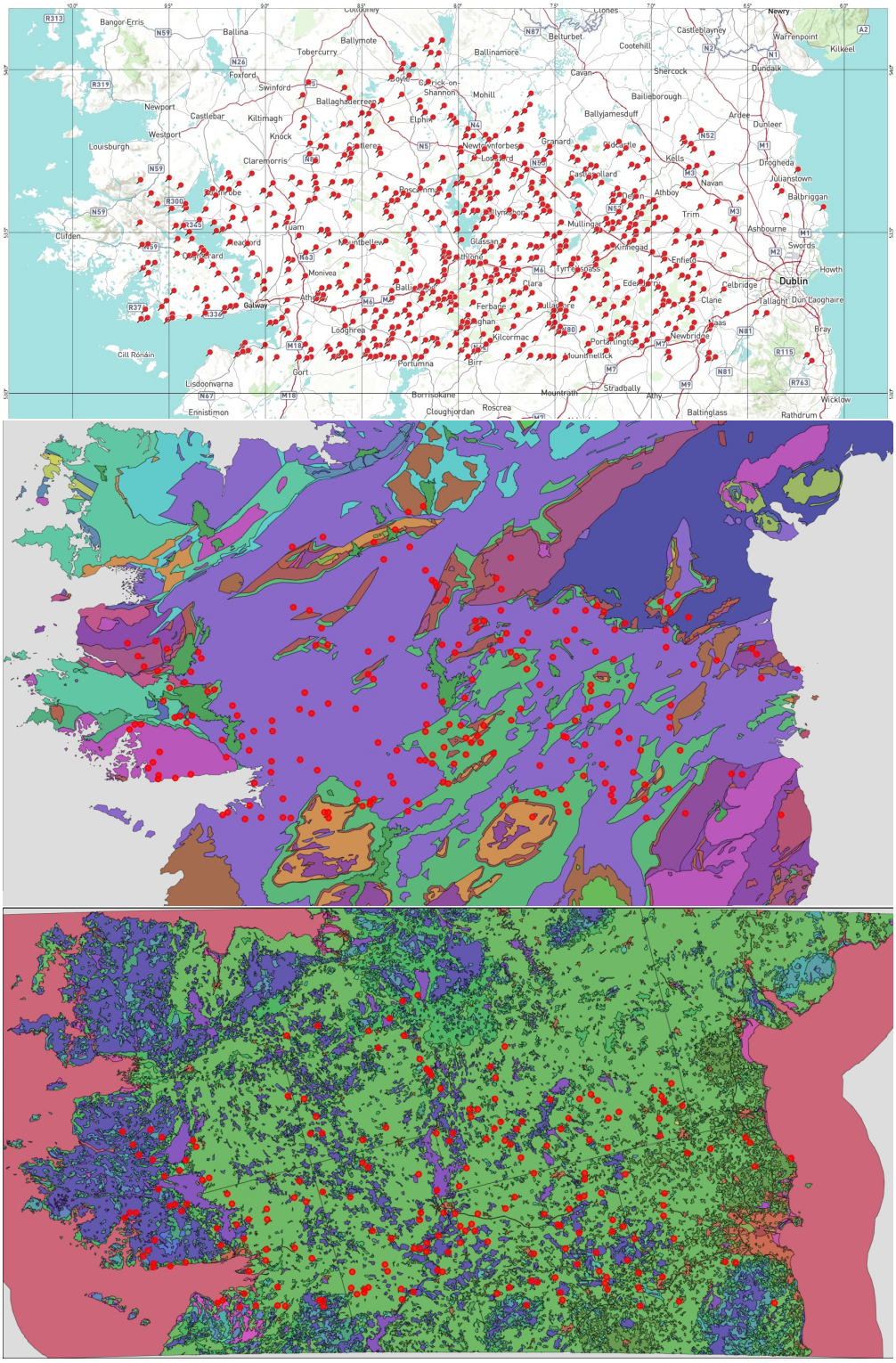


图 9: 异常集合中土壤样本的地理分布。上图：所有异常样本。中图：205 个高浓度异常值叠加在基岩地质上。下图：将相同的 205 个异常值映射到土地覆盖类型上。