FME 如何下“矢量数据融合”这盘棋

编　写：曹 文 涛

武汉市测绘研究院

二〇一九年三月

一、概述

随着经济社会的不断发展进步, 尤其是互联网地图等新型地理空间数据的出现与迅猛发展，各行业对于矢量地理空间数据的现势性和准确性要求越来越高，因此, 针对不同现势性、不同精度、不同来源的矢量数据融合技术进行研究具有十分重要的意义。这有利于降低地理空间数据的生产成本，加快其更新速度，提高其质量、精度和现势性，从而促进不同空间数据部门间的合作，推动空间数据的集成与共享，为城市社会发展和规划决策提供必要的技术和数据支持。

堪称数据处理界“瑞士军刀”的 FME，对于矢量数据融合这个研究热点来说，会有怎样的表现呢，这盘棋能否下得精彩呢？本文就利用 FME 对矢量数据融合进行一些探讨。

二、矢量数据融合

2.1 融合的条件与内容

首先，当待融合的数据源具有一些特点时，即前提条件,进行数据融合才有一定的意义，才能提高原始数据应用的深度、广度和增值程度。这些条件主要包括以下几个方面：

（1）在空间方面,待融合的数据源具有重叠区域；

（2）在尺度方面，待融合的数据源在尺度方面是相同或相近的；

（3）在时间方面，待融合的数据源是处于同一时期或者相近的时期；

（4）在数据特点方面，待融合的数据源应在精度、属性、现势性、详细程度、数据完整性等方面有所侧重。

当不同数据源符合以上若干条件时，才具备融合的价值和意义。

其次，由于各种矢量数据生产软件的功能、特点和优势等不同，造成矢量数据的格式、组织结构、表达方式等有较大的差异，因此在进行数据融合时，需要考虑空间数据模型、几何图形、实体属性、数据质量等方面的融合，在融合的过程中，一般分为数据预处理、同名实体匹配、几何图形调整和属性信息融合等内容，将在以下章节中进行具体的阐述。

2.2 数据预处理

数据预处理是指对不同的数据源进行规范化处理，使得它们具有相同的格式、坐标系统和地图投影等。在 FME 中，数据预处理中的格式统一问题可以忽略，它本来就能兼容读取各种不同格式的矢量数据，而空间位置的统一分为两种情况，如果待融合数据源的坐标系统和投影类型已集成在 FME 的坐标系统定义文件中，则可直接调用 Reprojector 转换器；如果未集成，则可通过向 LocalCoordSysDefs.fme 和 MyCoordSysDefs.fme 文件中添加坐标系的相关定义描述，然后再调用 Reprojector 转换器进行坐标转换即可，如图 1 所示，考虑到篇幅问题，其他预处理内容这里不再描述。

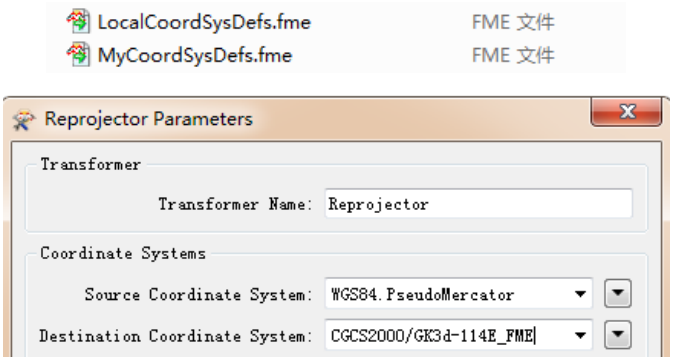


图 1 FME 中涉及坐标转换的配置文件与转换器

2.3 同名实体匹配

去年参加博客大赛时，写到过多尺度线和面要素的一致性检测内容，其中就涉及到同名实体的匹配，主要从空间位置、几何形状、拓扑关系和语义属性等四个方面进行考虑。设待匹配的要素 A 和 B，二者的要素相似度$\mathrm{ObjSim(A,B)}$是将空间相似度$\mathrm{S(A,B)}$、几何相似度$\mathrm{G(A,B)}$、拓扑相似度$\mathrm{T(A,B)}$和语义相似度$\mathrm{sem(A,B)}$等结合在一起，综合评价它们之间的整体相似情况，其值的取值区间为[0,1]，ω_S、ω_G、ω_T和ω_{sem}分别代表空间、几何、拓扑相似度和语义相似度在评价要素相似度时的权重值，须满足$\omega_S+\omega_G+\omega_T+\omega_{sem}=1$，要素相似度$\mathrm{ObjSim(A,B)}$的计算公式如式（1）所示：

$$\mathrm{ObjSim(A,B)}=\mathrm{S(A,B)}*\omega_S+\mathrm{G(A,B)}*\omega_G+\mathrm{T(A,B)}*\omega_T+\mathrm{sem(A,B)}*\omega_{sem} \quad (1)$$

从 A 和 B 的候选集中选取要素相似度最大的匹配组合，如果其要素相似度

大于阈值，则表明这对组合即为最优的要素匹配结果，从而实现了不同数据源之间同名实体的匹配，由于这部分内容在去年的博客大赛参赛文中作了详细的介绍，本文就不再赘述。

2.4 几何图形调整

几何图形调整是在同名实体匹配的基础上，将不同数据源的同名实体空间位置进行的调整，它不仅需要调整已匹配实体的位置，还需要利用已匹配实体去调整未匹配实体的空间位置。首先提到一个“可信度”的概念，它是指矢量数据中要素位置接近“真实位置”的程度。不同数据源间同名实体的位置对融合后的结果图形有不同的影响，因此需要对每个要素进行一个总体上量化的影响度衡量，以确定该要素在不同数据源中的可信度因子。本文选择同名实体对周围地物的影响度、位置的准确度和在不同数据源中的重要性这三个评价因素来决定其位置的可信度。

（1）影响度

本文利用 Delaunay 三角网来确定各要素对周围地物的影响度，矢量数据中的实体类型分为点、线和面三种，其中线可以理解为由多个有序的点组合而成，而面则是由多条线组合而成，通俗来说，要素对周围地物的影响度是指构成要素的点或点集合与构成其周围要素的点或点集一起构建 Delaunay 三角网，约束条件为要素本身，在约束三角网中与要素本身的点或点集合相关联的所有三角形面积之和即为要素对周围地物的影响度，如下图 2 所示：

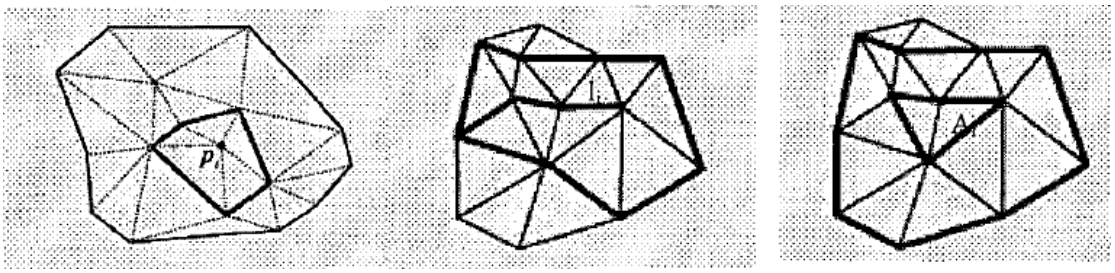


图 2 点、线和面对周围地物的影响度

那么在 FME 中，实现影响度计算的过程如图 3 所示：

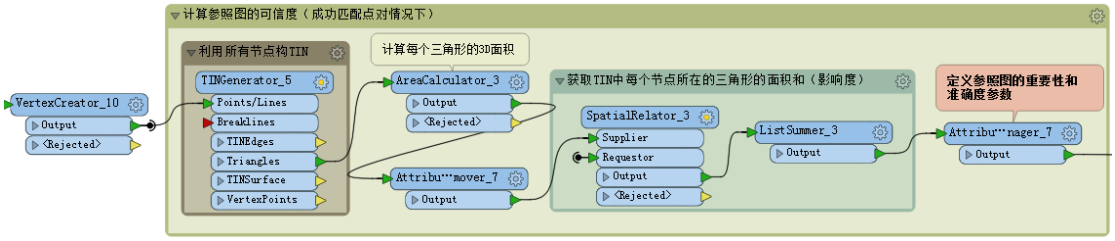


图 3 FME 中实现影响度计算的过程

通过 CoordinateExtractor 或者 VertexCreator 转换器获取要素与其周围要素的节点，然后利用 TINGenerator 转换器构造 TIN，并通过空间分析转换器 SpatialRelator 转换器获取与要素节点相关的三角形，最后利用 AreaCalculator 求其面积和，即影响度。

（2）准确度

不同数据源有不同的精度指标，一般来说，平面位置中误差M是评价数据精度的常用值，要素位置准确度与中误差成反比，误差越小，则该数据源中要素位置越接近于其真实位置，因此位置的准确度在数值上可用中误差的倒数$1/M$表示。

（3）重要性

不同数据源的重要性是有差异的，往往需要根据所要融合后的图形应用领域进行评价，这里重要性程度主要依据经验值确定。

（4）图形调整变换算法

在矢量数据中，面要素的表达最为复杂，且可以认为其是由线或者点构成的，所以这里以面要素的调整变换算法为例，来具体阐述几何图形的调整过程。

如图 4 所示为面要素的图形调整变换思路：

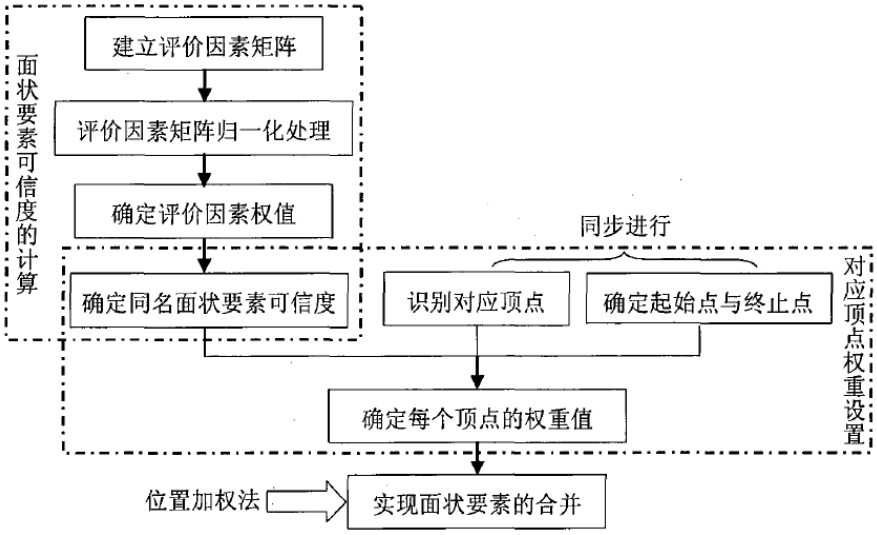


图 4 面要素的图形调整变换思路

2.4.1 评价因素矩阵建立及其归一化

面要素的三个评价因素，即影响度、准确度和重要性，它们的计算可按照前文所述实现，由于不同的评价因素量化值的数量级不同，所以要对其进行归一化处理, 否则评价因素之间的比较就没有任何意义。如图 5 所示：

Attributes To Create

New Attribute	Attribute Value
影响度因子_参照图	@Value(影响度._sum)/(@Value(影响度._sum)+@Value(调整图影响度._sum))
影响度因子_调整图	@Value(调整图影响度._sum)/(@Value(影响度._sum)+@Value(调整图影响度._sum))
准确度因子_参照图	@Value(准确度)/(@Value(准确度)+@Value(调整图准确度))
准确度因子_调整图	@Value(调整图准确度)/(@Value(准确度)+@Value(调整图准确度))
重要性因子_参照图	@Value(重要性)/(@Value(重要性)+@Value(调整图重要性))
重要性因子_调整图	@Value(调整图重要性)/(@Value(重要性)+@Value(调整图重要性))

图 5 评价因素矩阵的归一化

2.4.2 评价因素权值的确定

本文按照信息论中的熵法来确定权重, 表明各评价因素对面要素位置的重要程度。在归一化后的评价因素矩阵$\boldsymbol{G}=(g_{\mathrm{ij}})_{\boldsymbol{n}*\boldsymbol{m}}$中，$g_{ij}$出现的概率可由式（2）计算：

$$p_{ij}=\frac{g_{ij}}{\sum_{i=1}^{n} g_{ij}},\forall i,j \tag{2}$$

那么评价因素 j 的熵S_j和权重f_j分别由式（3）和（4）计算：

$$S_j=-\sum_{i=1}^{n} p_{ij}\ln p_{ij},\forall j \tag{3}$$

$$f_j=\frac{1-S_j}{\sum_{j=1}^{m}(1-S_j)},\forall j \tag{4}$$

如图 6 和 7 所示为评价因素熵和权值的计算。

Attributes To Create

New Attribute	Attribute Value
影响度_熵	-@Value(影响度因子_参照图)*@log(@Value(影响度因子_参照图))-...
准确度_熵	@Evaluate(-@Value(准确度因子_参照图)*@log(@Value(准确度因...
重要性_熵	-@Value(重要性因子_参照图)*@log(@Value(重要性因子_参照图))-...

图 6 评价因素熵的确定

Attributes To Create

New Attribute	Attribute Value
影响度_权重	(1-@Value(影响度_熵))/(1-@Value(影响度_熵)+1-@Value(准确...
准确度_权重	(1-@Value(准确度_熵))/(1-@Value(影响度_熵)+1-@Value(准确...
重要性_权重	(1-@Value(重要性_熵))/(1-@Value(影响度_熵)+1-@Value(准确...

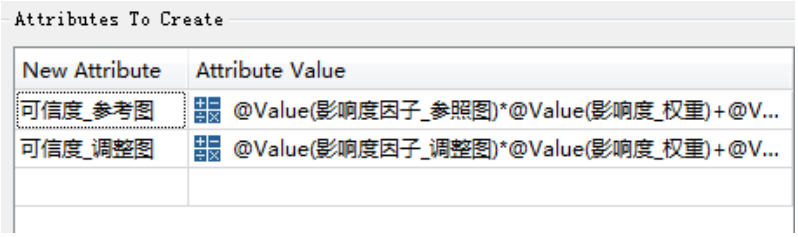
图 7 评价因素权值的确定

2.4.3 同名要素可信度的确定

同名面要素的可信度与评价因素以及其权重是相关的，可由式（5）计算：

$$\omega_j = \sum_{j=1}^{m} g_{ij} f_j, \forall i \tag{5}$$

如图 8 所示为不同数据源同名面要素的可信度计算：



Attributes To Create

New Attribute	Attribute Value
可信度_参考图	@Value(影响度因子_参照图)*@Value(影响度_权重)+@V...
可信度_调整图	@Value(影响度因子_调整图)*@Value(影响度_权重)+@V...

图 8 同名面要素可信度的确定

在同名面要素可信度的确定过程中，涉及到了评价因素归一化、权值确定、熵的计算和权重值的确定，这些均是一些数学公式的计算过程，直接用若干个 AttributeCreator 就可以搞定，无需其他的转换器。

2.4.5 成功匹配的对应点对调整变换

为了获取高质量的同名面要素融合后的结果图形，同名面要素上对应节点的识别是一个非常关键的步骤。本文利用离散 Fréchet 距离这个概念，它可以用来度量两线之间的最短距离，而面要素是由线构成的，故可以通过离散 Fréchet 距离来确定同名面要素上的对应节点对。

如图 9 所示在 FME 中利用离散 Fréchet 距离来确定同名面要素上的对应节点对的整个过程。

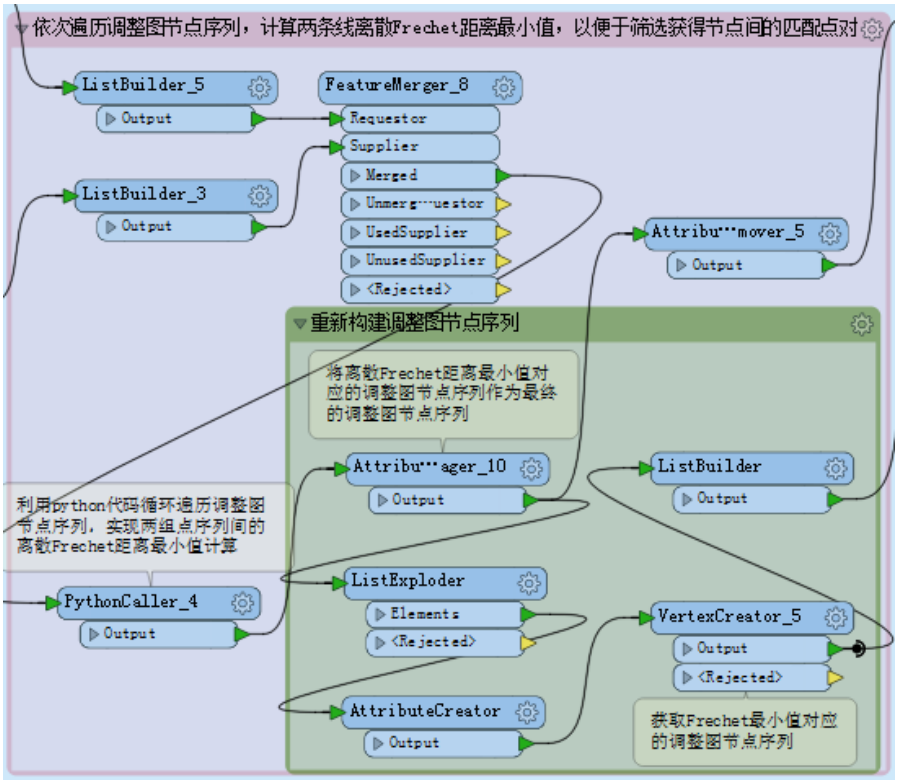


图 9 利用离散 Fréchet 距离确定对应节点对

由于离散 Fréchet 距离的计算步骤较为繁琐，这里不再具体列出，通过它来确定对应节点对，要利用 CoordinateExtractor 转换器获取同名面要素的所有节点，即 list 列表，由于面要素的起始节点并不确定，为获取面要素有序点集，须以每一个节点为起始节点循环重构节点序列，计算不同起始节点情况下同名实体间的离散 Fréchet 距离，取距离最小值作为同名面要素间的离散 Fréchet 距离，即最佳节点配对路线，并以当时的节点序列作为同名面要素的节点顺序，进行同名面要素节点对的融合变换。如图 10 所示为在 PythonCaller 转换器中，循环重构节点序列确定并输出最佳节点配对路线代码：

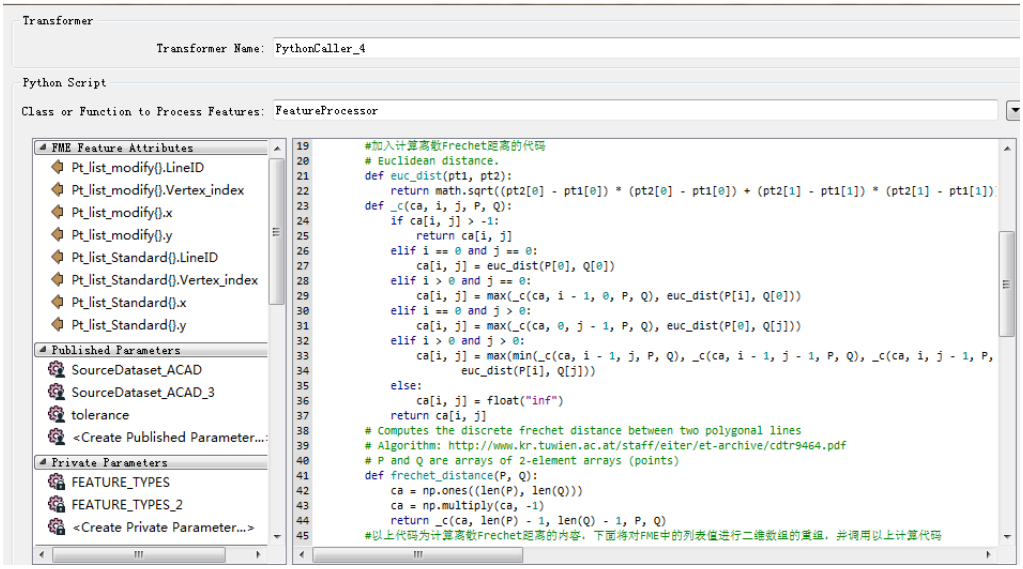


图 10 利用 PythonCaller 确定最佳节点配对路线

确定同名面要素的对应节点对后，还需要考虑节点所在的面要素的可信度和最佳节点配对路线中对应节点出现在所有节点对中的次数，这里主要将图 10 输出的最佳节点配对路线上的点对集合进行简单的图属关联和元素个数计算，主要用了 ListElementCounter 和 FeatureMerger，如图 11 所示：

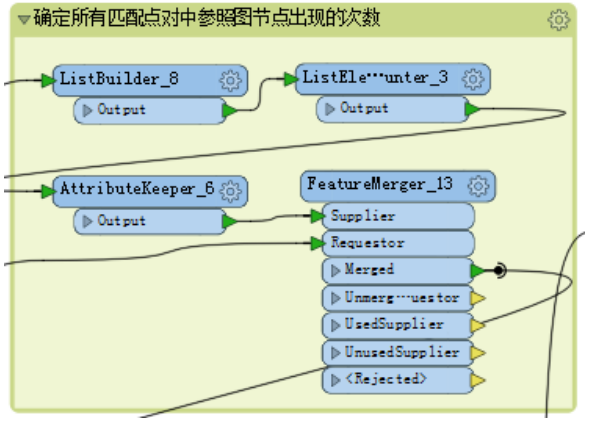


图 11 确定面要素节点在所有对应点对中出现的次数

次数的倒数与可信度的乘积即为不同数据源中对应节点对的权重值，再结合对应节点对的坐标值，进行加权平均，如图 12 所示，计算融合后成果图形的节点位置。

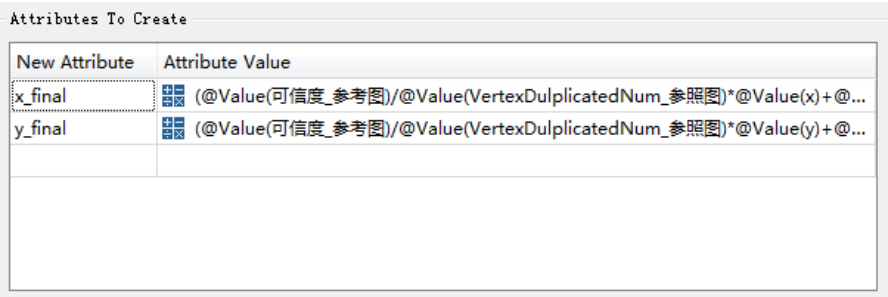


图 12 计算融合后成果图形的节点位置

2.4.6 未成功匹配点的调整变换

以上讨论的是基于成功匹配的对应点对的调整变换，对于未成功匹配的点，本文采用基于拓扑关系的形状校正调整变换算法，能够实现非规则图形在调整变换前后尽量保持其形状一致。首先计算所有成功匹配的节点对未匹配点的坐标总调整量，主要用到了 StatisticsCalculator 进行求和统计；然后按照调整前后同名实体上未成功匹配的点应与和它最近的两个已匹配同名点保持夹角不变为原则，计算调整后的位置，这里用到了 NeighborPairFinder 转换器；最后对这两个坐标值进行加权平均，即可得到未成功匹配点在调整变换后的位置。如图 13 所示：

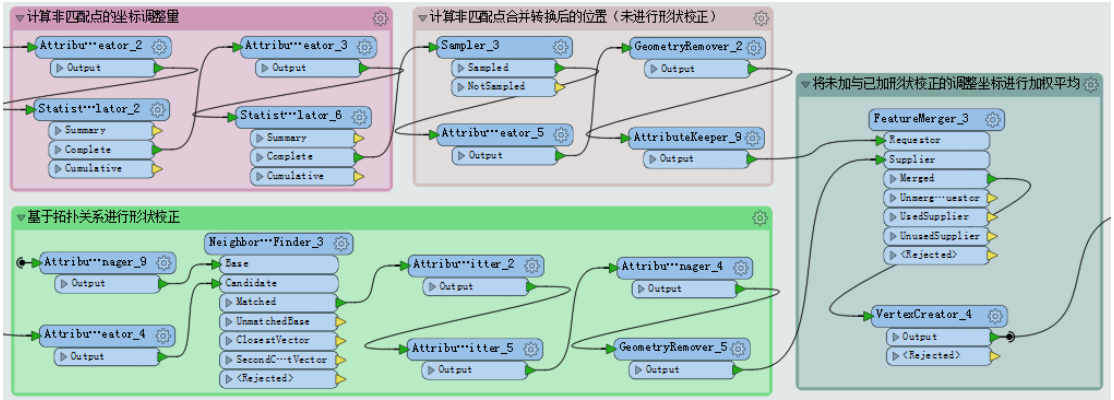


图 13 未成功匹配点的位置调整变换

综上，可以实现同名实体的几何图形位置调整，获得融合后图形的点集合，包括成功匹配点对的融合后点集（以下简称 S1）与未成功匹配点调整变换后的点集（以下简称 S2），后者的出现是随机无序的，因此对于线和面要素来说，如何在 FME 中实现通过点集合有序构建线和面要素呢？本文以类似“贪吃蛇”的思路，首先将有序的 S1 构建成线或面要素（以下简称 G）；然后遍历 S2 中所有节点，获取每个节点与 G 距离最近的一段线，然后将该节点融入这段线，直至

S2 为空集，结束循环，将该功能封装为自定义转换器“RetroPoints” ，如图 14 所示，最终可以得到同名实体经过几何图形调整后的成果图形。

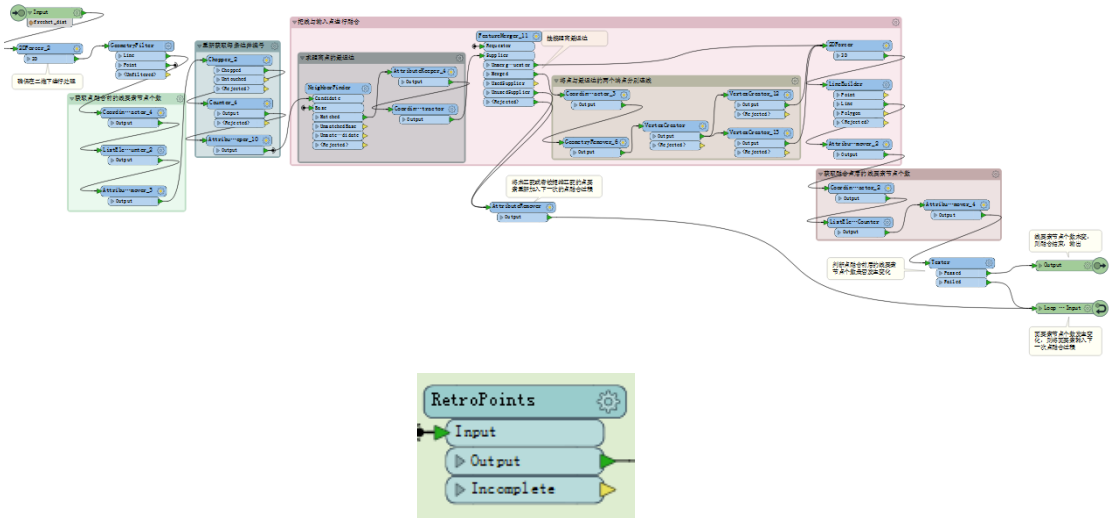


图 14 RetroPoints 自定义转换器

2.5 属性信息融合

对于属性信息的融合，原则上以“丰富”为准，可以将一种数据源对某种实体的属性描述直接扩充到另一种数据源对该实体的属性描述上，在 FME 中有很多转换器含有“Attribute Accumulation”的设置项，例如 FeatureMerger、Overlayer 系列、Spatial 系列等，完全可以根据用户的需求来定制属性的操作；有时也可以根据具体应用需要，对属性项进行抽象、概括和重构，那么利用最简单的 AttributeCreator 就能够搞定。

三、融合效果

在本文第二部分内容中，对矢量数据融合的每个环节做了具体的阐述，基于以上的讨论，在 FME 中编写了矢量数据融合的模板，如图 15-17 所示：

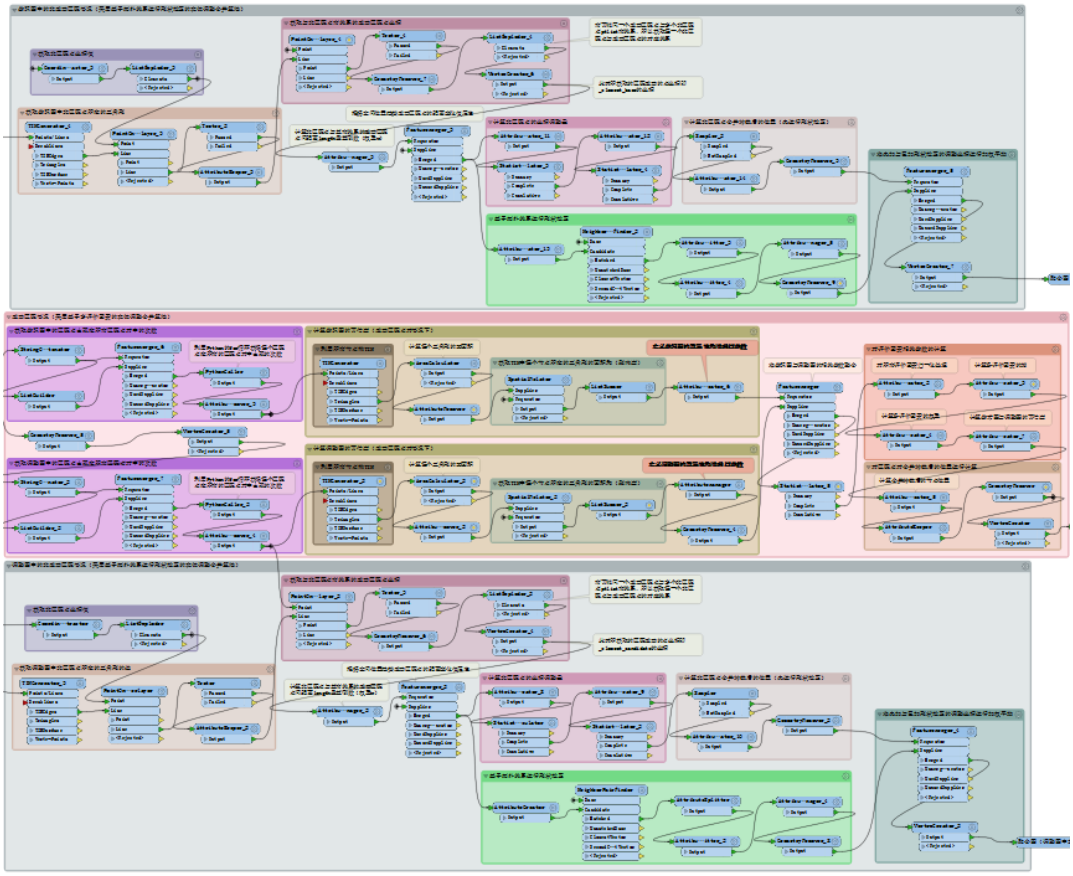


图 15 点要素融合模板

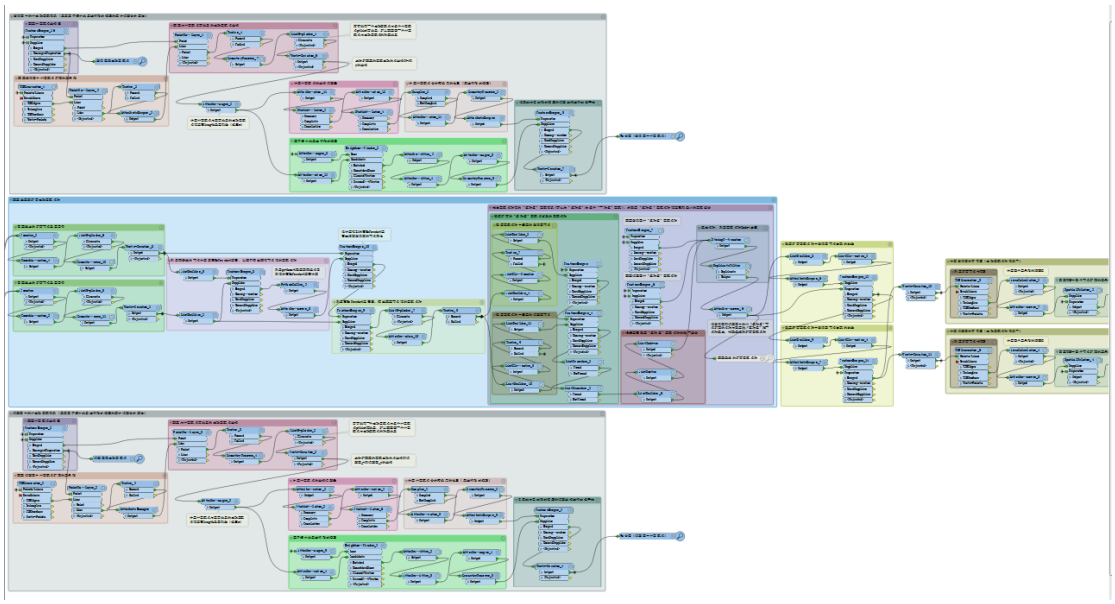


图 16 线要素融合模板

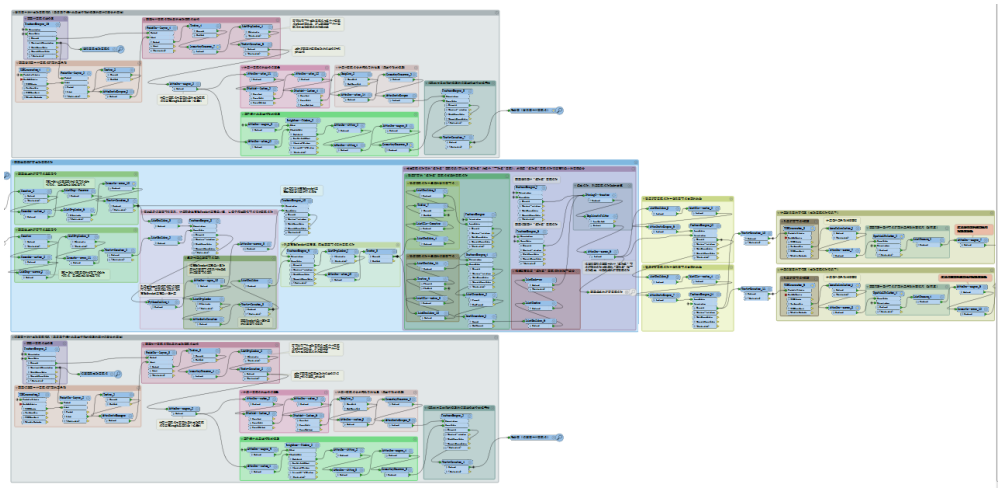


图 17 面要素融合模板

那么点、线和面要素的融合效果如图 18-20 所示：

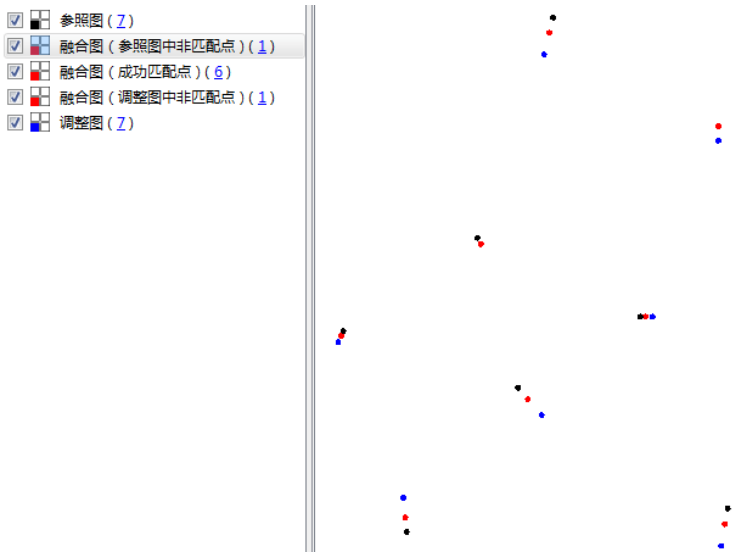


图 18 点要素融合示例

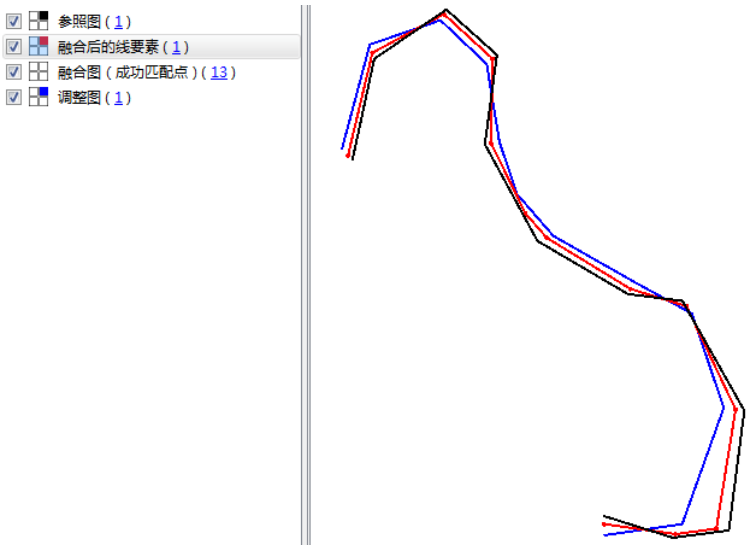


图 19 线要素融合示例

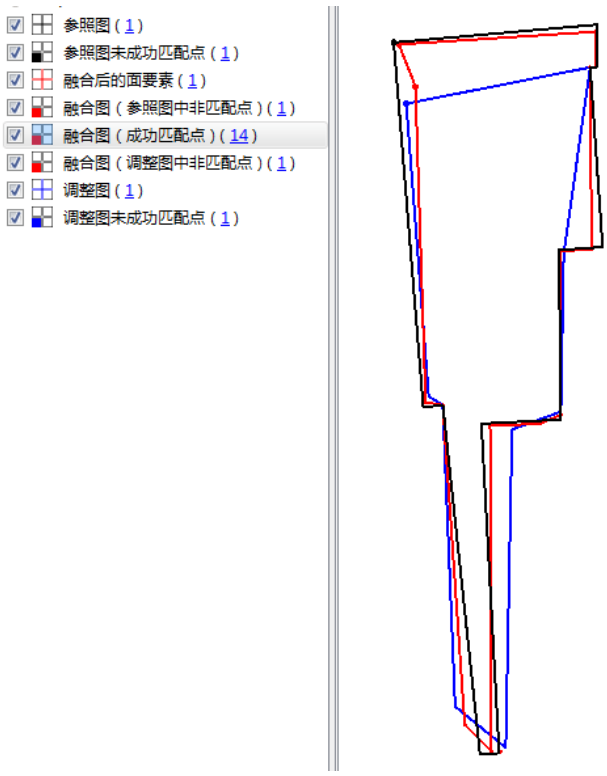


图 20 面要素融合示例

以上点、线和面的融合效果较好，从而验证了本文所阐述矢量数据融合方法和思路的有效性和可用性，这说明用 FME 来下“矢量数据融合”这盘棋是可行的。矢量数据融合的成果可以根据不同的融合需求，达到不同的融合效果，生成不同特点的数据成果，这就降低了数据的生产成本，大大提高了已有数据源的应用领域和价值。

四、结语

对于矢量数据融合的研究，仍然是目前地理信息行业的热点，它是智慧城市建设过程中不可或缺的技术手段。本文将矢量数据融合的主要内容、融合过程和思路做了一些探讨，获得了较好的矢量数据融合效果。笔者由于能力有限，文中很多地方可能有错误和不足，欢迎有兴趣的同志们指导和交流。