**图解税收大数据分析**

叶健峰 著



卷首语

终南别业

唐 王维

中岁颇好道 晚家南山陲

兴来每独往 胜事空自知

行到水穷处 坐看云起时

偶然值林叟 谈笑无还期

前言

整理一下这几年在税收数据分析上的学习研究，梳理一下体系条理，以后慢慢补上缺失的部分，还有神经网络与深度学习未曾做出NLP之外好的税收分析实例。从这些应用场景可以看到，税收大数据分析对云平台支持的要求是多元的，包括了关系数据库、图数据库、NLP、GIS、机器学习、深度学习、API、GPU、可视化、数据集成等方面，在所有方面都达到落地应用要求已属不易，都达到行业领先更是比较困难，对此应该有全面的了解，心中有数，不必责备求全。随着相关平台的研发建设，博采众人之所长，集成整合，目标总是可以达到的，有些事情或者只是早一点晚一点的问题。

整理成一份有体系有条理的文档是有必要的，可能还没有达到出版成书的水平与篇幅，但有助于看清大局，梳理出可行的技术路线和需要努力的方向。

个税与社保的分析文章几年前就已有文稿，这几年陆续学习研究了发票分析的NLP货劳名称识别、图数据库交易网络分析、风险模式识别与机器学习、深度学习、时间序列分析、GIS融合应用、文本挖掘等落地应用专题，又积累了一些文稿。写着写着，好像也有点模样了。从数据量上看，社保、个税、发票是数据量最大的，都深入研究了；从业务上看，企业所得税、增值税、土地增值税是最复杂的，还未有时间机会去涉猎与深入。其它具体税种的分析文章，以后有机会慢慢学习，慢慢补上。

本书收录了这几年发在自媒体上的系列学习研究文章，适应自媒体手机传播阅读的特点，文字力求简洁，图表力求精要，以图片为主。程序代码等原来在手机上展示阅读不太方便，所以以截图为主，手机上放大阅读则比较方便。结集成Word电子书后，图片在页面中都有不同程度的压缩，在电脑上阅读时，可以逐张把图片拷贝另存出来全屏看，程序代码则恢复为文本以便动手尝试。

这些实例包括了R语言、Python、Java、SQL、Cypher、JavaScript、JSP、HTML/CSS等多种编程语言，随着我的学习研究按需要使用，业务与逻辑是通用的，很容易就可以在不同的语言和平台间移植，因精力有限，结集成书时还是保持原来的样子，具体语言与平台的细节就不在本书中展开介绍。

这些实例并非全部是税收分析实例，比如图算法增强机器学习的连接预测与节点分类，时间序列分析的季节ARIMA及GARCH模型，文本挖掘的微信、微博分析，深度学习的气温预测与照片分类，但它们显然在税收大数据分析的落地应用中具有很大的潜在价值，覆盖了税收大数据分析的主要方面，可以参考。所以作为技术与概念介绍的实例收录书中，这也是我学习的一个过程。

这些实例的学习与研究，历时数年，如果从2011年前后从SAS学习数据挖掘开始，大约也有10年的积累了。所以这本书也可以说是写了几年，是一个阶段的结束，整理成书，算是对自己的一个交代。笨鸟先飞，普通人只能象曾国藩读书一样，用苦功夫、笨功夫。没有能力造飞机，咱就努力尝试造副适用于税务机场的起落架吧。

另外，根据单位相关部门的建议，按照税总[2021]29号文的要求，所有发在自媒体上包含涉税数据的文章，都已收回，结集成本书。

最后，各省基本上都已实现了省级集中，这些年税收信息化的实践表明，下一阶段必须有统一的顶层规划与实施，不管是省级还是国家级。但税收大数据应用的顶层规划与落地需要一个较长的周期，所以在附录中也介绍了中小型的开源过渡解决方案Rstudio Linux Server等，以解决基层眼前的一些急需。

自序

我己经50岁了，知命之年，却道天凉好个秋。从癌症中康复过来，洗尽浮华，归根复命，一把年纪付出巨大的努力投身于学习研究税收大数据分析，已逾8年。孤军奋战能够坚持下来，是因为我相信就全国全税务系统来说，要改变局面，建立有效的体系，纲举目张，需要这样做。从死亡边缘回来的人，生命有限，能量有限，万事归空，生命中已经没有不能承受之轻，也不会去追求升官发财了，那就安安静静的做点有价值有意义的事吧，力所能及的事就是自己熟悉的事了。我不想浪费时间去敷衍不相干的人和事，这样的状态，从心所欲不逾矩最好。功成不必在我，功成必定有我，倘若我的研究和付出能对国家社会有所贡献，那是值得欣慰的。还有不到2年就工作满30年，那时大概可以考虑提早退休了吧。所以我不一定能够在退休前看到研究的成果付诸应用了，于是有了把研究文章结集成书的想法。结集成书，传播火种，既是对我这8年付出努力的一个交代，也是这一段生命留下的印迹。

目前税收大数据的主要应用仍是宏观领域的分析应用，微观领域主要是行业与个案的案头手工分析，缺乏当代大数据技术下微观税收分析的体系与实例，这是我近期搜索已出版书籍的结果。税收大数据要发挥出它的价值，必须深入微观分析的领域，这是不应回避也无法回避的。

税务系统需要建立当代大数据技术下完整的微观税收分析体系，贯通宏观与微观，集成大数据、图计算、NLP、GIS、机器学习、深度学习等各种现代信息技术与数据科学，宏观税收分析及经济分析才能基础牢固有突破，微观风险管理才能更精准，纳税服务才能更精细贴心，内部管理才能更高效，才能达到征管改革合成的目标。

税务系统已经积累了海量的数据，但非常欠缺AI机器学习所需的大量标签数据，不解决从0到1的问题，海量数据的价值就难以充分发挥，风控水平也难上一个台阶。技术层面需要解决发票分析的交易对象确定、交易网络分析、交易对象分类（商品税收分类代码映射）算法、虚开网络建模等关键问题，各种微观分析主题中快速数据标签算法的问题等。这些都是税收大数据平台必须解决的基础性问题，我的研究对此都进行了探索，尝试打通税收大数据应用最后的一公里。它们并没有理论上的创新，属于落地应用上的拓展，解决的是应用中的工程技术问题，也许可以算是应用创新吧。

在具体的微观层面，税收问题都是涉及几十万乃至几亿以上的问题。税收政策不断的新出或打补丁，每年的税收文件雪花一样源源不绝，因此会有各种各样解释和操作的空间，里面可能会有一些不得以不方便不足以不能够为外人道的事情，它们大概会成为大数据微观税收分析落地应用的主要阻力。我研究的主要方向是机器学习，用算法对历史数据作全量的扫描与挖掘，例如个税年薪连年下降风险识别，二手房转让价格回归分析，X光一样，所有问题无所遁形，是大规模杀伤性武器，所以阻力之大难以想象。国之大者，当胸怀天下，我相信就全国全税务系统来说，要改变局面，建立有效的体系，纲举目张，就需要深入微观领域。一方面顶层规划与国家意志必不可少，以打破一些固有的利益格局，强力推进，《关于进一步深化税收征管改革的意见》由中办国办联合发文已经体现了这一点。另一方面要谨慎使用、受控使用，理性宽容的处理历史上一些不够规范的问题，以消除各方面的担忧，减少阻力。数据分析的责任是廓清现状，找到问题，给出建议，如何处理是领导们商量的事。

个税大数据分析是我搞得比较透的一个落地应用主题，国地税合并，个税改革前就已完成了。自底向上对个税明细申报数据进行了全面、深度的分析，贯通覆盖了风控、税政、征管、收核、技术等各主要业务领域，展示了税收大数据应用在数据抽取与整合、微观宏观分析、AI建模等方面实际应用中灵活多变、交互、迭代的特点。业务与技术的深度融合，对业务、技术人员提出了更高的要求，也提出了对大数据平台提供灵活支持、可编程支持及API支持的要求。这个完整的深度落地应用实例，希望能为相关大数据平台规划及业务需求梳理提供一些参考。大数据应用平台要解决什么问题，应该怎样规划，应用应该怎样开发，流程应该怎样组织，人员团队应该怎样组织培养，组织机构及职能怎样划分调整，怎样用好大小厂等社会力量，通过实现这样一个有深度、全面、完整的实例，希望各级读者会有更多的思考。

从这些实例，可以看到各种语言和工具怎样配合运用去完成分析任务，实现业务与技术的深度融合，可以看到编程是深度分析唯一途径。这正是世界上大多数数据科学工作者在做的事。税收治理是国之重器，国家应确保掌握和主导，所以税务系统首先要培养起自己业务与技术相融合的复合型人才队伍。

书中的实例，覆盖了微观税收分析中数据探索、分析、建模、模型发布、撰写分析报告等的完整工作流程，也展示了开源生态对上述各数据应用环节的技术支持。因为是个人学习研究，资源有限，基层小人物也不适合卷入大厂之间激烈的竞争，所以我选择了Rstudio Server、Shiny Server及Spyder、Jupyter Hub、Orange等开源数据科学工具以及HanLP、Neo4j等开源产品，云端环境则测试了本省税务系统都有使用的阿里云与华为云。实际上，这些开源工具和产品也有很强的连接能力，连接各大厂等各种国产平台也完全没有问题。相关的业务分析、业务逻辑与模型是通用的，也很容易移植到各种平台上。我个人的一点实践经验与观点，水平有限，人微言轻，无伤大雅。万物并育而不相害，这是大厂们应有的胸襟。

我只是想安安静静的做一点力所能及的事，于国于家有益，活得有一点尊严。上善若水，水善利万物而不争。处众人之所恶，故几于道。

此心光明，夫复何言。是为序。