**交通雾霾相关性分析小组工作进展报告(20160405)**

胡煜成，姚嘉豪

**目标：**交通数据整理，相关性分析方法调研

**阶段进展：**

1. 建立数据库

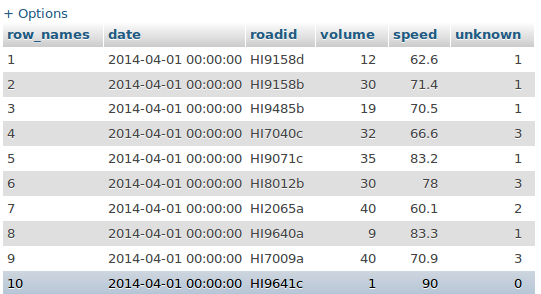
搭建了MySQL数据库。

数据库地址: hydrays.synology.me

用户名: user (拥有可添加数据权限)

密码: traffic2016(对外宣传时此处请隐藏)

数据库名称: [mwTrafficData](http://hydrays.synology.me/phpMyAdmin/db_structure.php?server=1&db=mwTrafficData&token=b9378f5c2bade7ce2bdf5d2249c9eef0)，内含“traffic”表单，存储了2014.01.01~2014.06.30共六个月的路段微波数据，字段名目前为”date”(时间戳), “roadid”(路段ID)，“speed”(路段平均速度), “volume”(路段平均流量？), “unknown”(未知暂时不清楚其含义)。数据样本截图如下：



2. 通过R访问数据库

R里面可以通过RMySQL包访问MySQL数据库(有待补充)。

# 载入包

require(RMySQL)

# 输入用户密码信息

con <- dbConnect(dbDriver("MySQL"), dbname = "mwTrafficData",

user="user", password="\*\*\*\*\*\*\*\*", host="hydrays.synology.me")

# 数据列表

dbListTables(con)

# 获取数据，可以带条件，数据量很大，时间会比较长(有什么提速办法吗？)

z <- dbGetQuery(con, paste("select row\_names, speed from traffic",

"where volume > 100 order by speed"))

# 退出

on.exit(dbDisconnect(con))

具体请参考R帮助文件。

3. 相关性分析方法调研

经过前期初步分析发现，交通可雾霾的关系不容易通过数据反应出来，因此需要深入考虑，适当运用统计工具，几种可能的方法如下(有待补充)：

(a) Granger-causality: 分析两个时间序列的因果关系，在宏观经济学里面比较成功。基本思想是：对于时间序列X(交通)和Y(雾霾)，如果运用X和Y的历史预测Y比仅用Y的历史预测Y要更准确，则表明X可能对Y有影响。

(b) ConvergentCross Mapping: 对于非线性动力系统中X和Y的关系的描述，是一个很有创意的想法，短时间内难以实现，长远来说值得借鉴，具体见参考文献：Sugihara, George; et al. (26 October 2012). ["Detecting Causality in Complex Ecosystems"](http://www.uvm.edu/~cdanfort/csc-reading-group/sugihara-causality-science-2012.pdf) (PDF). Science **338** (6106): 496–500. Retrieved 5 July 2013.

(c) 仿照Yu Zheng的2013年文献“U-Air: When Urban Air Quality Inference Meets Big Data”中的Figure6, 7进行画图。

**未来工作：**