汽车新产品竞争网络数据处理说明

数据处理的目的：基于汽车新产品数据构建厂商之间的竞争网络及其他相关变量，并将其导入汽车产业WTO面板数据

数据处理的原则

1. 清洗汽车新产品企业名称。汽车新产品的“整车数据”表格中ZZCMC字段表示该汽车产品的车辆制造企业。该字段下面的车辆制造企业名称存在不规范、不标准、不统一的情况，需对原始数据进行清洗，使得同一家公司的名称保持一致。对ZZCMC字段下的企业名称进行相似度分析，将高度相似的企业名称提取出来，结合“SCDZ”（生产地址）下的信息进行分析。当某两个车辆制造企业名称不完全相同，但高度相似，且SCDZ（生产地址）下的信息为同一城市，则判断该车辆制造企业为同一家企业，取最新出现的企业名称作为同同一后的企业名称。对于其他高度相似的车辆制造企业名称，则进行手动核查，结合企查查、天眼查等平台确定其是否为同一家企业。输出ZZCMC名称清洗表格参考如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ZZCMC名称A | 生产地址A | ZZCMC名称B | 生产地址B | 相似度 | 统一后ZZCMC企业名称 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

前述名称修改表格确定后，按照该数据表确定的原则，对原始“整车数据”表格中ZZCMC字段下的企业名称进行清洗。

1. 前述ZZCMC字段清洗完成后，将该字段下的企业名称与“AmIndustryWTOPanel20181128”中的OGNM下面的企业名称进行清洗配对，使得同一家企业在ZZCMC和“AmIndustryWTOPanel20181128”数据表的OGMN字段下的名称保持一致。当ZZCMC和OGNM字段下的企业名称完全一致，则判断其为同一家企业。当ZZCMC和OGNM字段下的企业名称高度相似，且SCDZ（生产地址）所在城市和AmIndustryWTOPanel20181128中的地址所在城市相同，则判断其为同一家企业。当ZZCMC和OGNM字段下的企业名称高度相似，但不属于前述情况，则进行手动核查，结合企查查、天眼查等平台确定其是否为同一家企业。其原则是保持AmIndustryWTOPanel20181128中OGNM下面的企业名称不变，对“整车数据”表格中ZZCMC字段进行清洗，使同一家企业在“整车数据”表格中ZZCMC字段和AmIndustryWTOPanel20181128中OGNM字段下保持一致。输出ZZCMC对OGNM企业名称清洗表格参考如下：

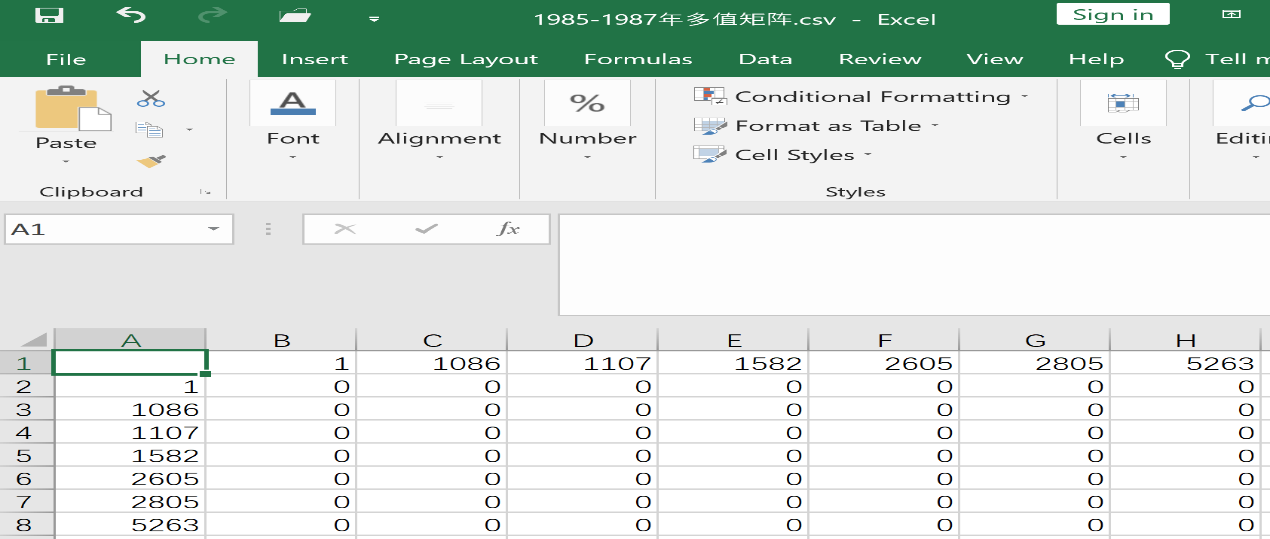
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ZZCMC名称A | 生产地址A | OGNM名称B | 生产地址B | 相似度 | 统一后ZZCMC企业名称 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

完成上述表格后，根据表格确定的原则对“整车数据”表格中ZZCMC字段下的企业名称进行清洗。清洗完成后制作整车数据ZZCMCZ编码表如下：

|  |  |  |
| --- | --- | --- |
| ZZCMC企业名称 | ZZCMC编号 | AmIndustryWTOPanel中IDNO |
|  |  |  |
|  |  |  |
|  |  |  |

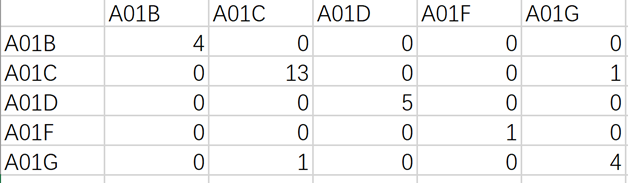
ZZCMC编号随机生成。

1. 基于前述两步清洗完成后的“整车数据”，构建3年窗口期的汽车新产品竞争网络。“整车数据”中CLLX（车辆类型）字段代表该产品的类型编码。当两家公司的产品属于同一个CLLX，则这两家企业存在竞争关系。根据这一原则构建企业之间的竞争网络矩阵，示意如下：



矩阵中的第一横行、第一竖行为企业的ZZCMC编号。任意两个ZZCMC编号对应的单元格中的取值代表这两家企业的竞争关系强度，等于这两个企业在某个3年的时间段内在“整车数据”表中有多少对产品属于同一个CLLX编码。例如，假设企业A、B在2008-2010年间有3对产品属于B1A、5对产品属于B23、10对产品属于B11，则A和B2008-2010窗口期竞争网络矩阵中对应的单元格取值为18。“整车数据”中一条数据对应一个产品。GGSXRQ（公告生效日期）视为该产品开始公开销售的时间。CXSXRQ（停止销售时间）为该产品停止在市场上销售的时间。这两个字段用来判断该产品在某个时间内是否在市场上销售。假设某产品公告生效日期为2000.01.01，而停止销售时间为2010.05.30。则2000-2010年间所涉及的窗口期中需考虑该产品是否与其他产品构成竞争关系。对于截至时间在2000年之前（比如1997-1999）的窗口期，则无需考察该产品是否与其他产品构成竞争关系；对于开始时间在2010之后（比如2011-2013）的窗口期，也无需考察该产品是否与其他产品构成竞争关系。对于停止销售时间缺失的数据，则默认其停止销售时间为2017.12.31。

1. 基于第3步的原则，构建5年窗口期的汽车新产品竞争网络。
2. 针对前述3年窗口期、5年窗口期的汽车新产品竞争网络，利用NETWORKX计算每个网络节点的下述网络指标，并以EXCEL格式保存其结果：degree centrality、shortest-path betweenness centrality、local reaching centrality, constraint(structural holes)、effective-size(structural holes)、clustering/triangles、clustering/clustering(the clustering coefficient for nodes)。NETWORKX计算网络指标的操作参见：https://networkx.github.io/documentation/stable/index.html。此外，计算每个节点的tie strength，即每个节点与它的竞争对手之间的链接强度的均值。比如A公司与B、C公司的竞争关系强度分别为2、5，则A当期的tie strength为（2+5）/2=3.5。
3. 在“AmIndustryWTOPanel20181128”数据表中新建如下变量，cpntdgr3yw/ cpntdgr5yw（3年/5年窗口期竞争网络degree centrality）、cpntbtw3yw/cpntbtw5yw（3年/5年窗口期竞争网络shortest-path betweenness centrality）、cpntlcrch3yw/ cpntlcrch5yw（3年/5年窗口期竞争网络local reaching centrality）、cpntcstr3yw/ cpntcstr5yw（3年/5年窗口期竞争网络constraint(structural holes)）、cpnteffsz3yw/ cpnteffsz5yw（3年/5年窗口期竞争网络effective-size(structural holes)、cpnttrgl3yw/ cpnttrgl5yw（3年/5年窗口期竞争网络clustering/triangles）、cpntclust3yw/ cpntclust5yw（3年/5年窗口期竞争网络clustering/clustering）、cpnttiestr3yw/ cpnttiestr5yw（3年/5年窗口期竞争网络tie strength）。然后把第5步计算出来的结果，填写到AmIndustryWTOPanel20181128面板数据表格的相应变量中去。
4. 针对“AmIndustryWTOPanel20181128”数据表中的企业，构建3年窗口期、5年窗口期IPCMG网络。IPCMG是指IPC分类号从第一位到“/”之前截至的所有编码。比如B60Q9/00的IPCMG是B60Q9。构建IPCMG网络只使用发明专利和实用新型专利中的专利分类号。当两个不同的IPCMG在同一个专利的分类号中同时出现，则认为这两个IPCMG之间存在共现关系。矩阵对角线上的取值为IPCMG自己与自己的共现次数，按当期在多少个专利中出现了IPCMG自己与自己共现的情况予以计算。分别以3年、5年为窗口期从某企业的所有发明和实用新型专利中提取IPCMG之间的共现关系，构建该企业的3年、5年窗口期IPCMG网络矩阵。采用的专利为在构建汽车产业面板数据中所使用的汽车产业专利数据。在文件夹中。IPCMG矩阵格式如下所示：

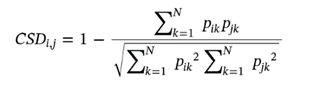


第一栏、第一行都为IPCMG编码，单元格中的取值为相应的两个IPCMG编码在3年窗口期或者5年窗口期该企业的所有专利中共现的次数。以EXCEL形式输出所有企业3年窗口期、5年窗口期IPCMG矩阵。如果某企业在某个3年或者5年窗口期内没有申请任何专利，则该企业在该窗口期内不存在IPCMG矩阵。如果某企业在某个窗口期的所有专利中存在IPCMG不与其他任何IPCMG发生共现关系，则该IPCMG在当期网络矩阵中为孤立的节点。

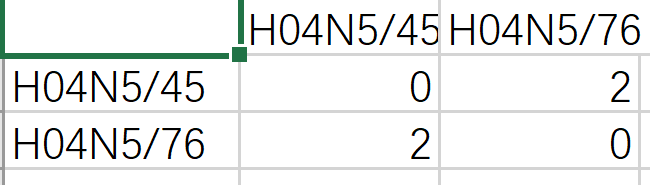
1. 针对“AmIndustryWTOPanel20181128”中的企业，利用NETWORKX计算每个企业每期IPCMG网络矩阵的degree centrality、shortest-path betweenness centrality、local reaching centrality, constraint(structural holes)、effective-size(structural holes)、clustering/triangles、clustering/clustering(the clustering coefficient for nodes)、tie strength的均值，以及network size（网络规模）、network density(网络密度)、clique（IPCMG网络中最大clique的规模）、efficiency（average global efficiency of the graph）、isolate（IPCMG网络中isolate的数量）。以EXCEL格式输出计算结果。使用NETWORKX计算网络指标的操作参见：https://networkx.github.io/documentation/stable/index.html
2. 在“AmIndustryWTOPanel20181128”中新建变量ipcmgdgr3yw/ipcmgdgr5yw(3年/5年窗口期IPCMG网络中的degree centrality均值)、ipcmgbtw3yw/ ipcmgbtw5yw（3年/5年窗口期IPCMG网络中的shortest-path betweenness centrality均值）、ipcmglcrch3yw/ ipcmglcrch5yw（3年/5年窗口期IPCMG网络中的local reaching centrality均值）、ipcmgcnstr3yw/ ipcmgcnstr5yw（3年/5年窗口期IPCMG网络中的constraint均值）、ipcmgeffsz3yw/ ipcmgeffsz5yw（3年/5年窗口期IPCMG网络中的effective-size均值）、ipcmgtrgl3yw/ipcmgtrgl5yw（3年/5年窗口期IPCMG网络中的clustering/triangle均值）、ipcmgclust3yw/ipcmgclust5yw（3年/5年窗口期IPCMG网络中的clustering/clustering均值） 、ipcmgtirstr3yw/ipcmgtirstr5yw（3年/5年窗口期IPCMG网络中的tie strength均值）、ipcmgntsz3yw/ ipcmgntsz5yw（3年/5年窗口期IPCMG网络规模）、ipcmgntdnst3yw/ ipcmgntdnst5yw（3年/5年窗口期IPCMG网络密度）、ipcmgcliq3yw/ipcmgcliq5yw（3年/5年窗口期IPCMG网络中最大clique的规模）、ipcmgeffc3yw/ipcmgeffc5yw（3年/5年窗口期IPCMG网络的average global efficiency）、ipcmgisolt3yw/ipcmgisolt5yw（3年/5年窗口期IPCMG网络中isolate的数量）。将第8步中的计算结果导入面板数据中相应变量的相应位置。
3. 针对汽车产业专利数据中的IPCSG编码，按照同样的规则重复第7、8、9步，针对“AmIndustryWTOPanel20181128”中的企业，构建3年窗口期、5年窗口期IPCSG网络，计算同样类型的网络指标，并导入面板数据。在面板数据中针对每个企业的IPCSG网络构建新变量：ipcsgdgr3yw/ipcsgdgr5yw(3年/5年窗口期IPCSG网络中的degree centrality均值)、ipcsgbtw3yw/ ipcsgbtw5yw（3年/5年窗口期IPCSG网络中的shortest-path betweenness centrality均值）、ipcsglcrch3yw/ ipcsglcrch5yw（3年/5年窗口期IPCSG网络中的local reaching centrality均值）、ipcsgcnstr3yw/ ipcsgcnstr5yw（3年/5年窗口期IPCSG网络中的constraint均值）、ipcsgeffsz3yw/ ipcsgeffsz5yw（3年/5年窗口期IPCSG网络中的effective-size均值）、ipcsgtrgl3yw/ipcsgtrgl5yw（3年/5年窗口期IPCSG网络中的clustering/triangle均值）、ipcsgclust3yw/ipcsgclust5yw（3年/5年窗口期IPCMG网络中的clustering/clustering均值）、 ipcsgtirstr3yw/ipcsgtirstr5yw（3年/5年窗口期IPCSG网络中的tie strength均值）、ipcsgntsz3yw/ ipcsgntsz5yw（3年/5年窗口期IPCSG网络规模）、ipcsgntdnst3yw/ ipcsgntdnst5yw（3年/5年窗口期IPCSG网络密度）、ipcsgcliq3yw/ipcsgcliq5yw（3年/5年窗口期IPCSG网络中最大clique的规模）、ipcsgeffc3yw/ipcsgeffc5yw（3年/5年窗口期IPCSG网络的average global efficiency）、ipcsgisolt3yw/ipcsgisolt5yw（3年/5年窗口期IPCSG网络中isolate的数量）。将相应的计算结果导入面板数据。
4. 将汽车新产品竞争网络中的企业名称与汽车产业专利数据中的企业名称进行匹配清洗，使得同一家企业在汽车新产品竞争网络中的名称与专利数据中的名称保持一致。采用的专利为在构建汽车产业面板数据中所使用的汽车产业专利数据。基本原则是保持汽车产业面板数据中所使用的汽车产业专利数据中的企业名称不变（因为在构建汽车产业面板数据过程，已经对专利数据的企业名称完成了清洗工作），保持本说明第1、2、3步已经清洗完成的企业名称不变，对汽车新产品竞争网络中剩下的企业名称进行清洗，使其与专利数据中的企业名称保持一致。当汽车新产品竞争网络中的企业名称和专利数据中的企业名称完全一致，则判断其为同一家企业。当汽车新产品竞争网络中的企业名称和专利数据中的企业名称高度相似，且汽车新产品竞争网络中的该企业的SCDZ（生产地址）所在城市和专利数据中该企业的地址所在城市相同，则判断其为同一家企业。当两个企业名称高度相似，但不属于前述情况，则进行手动核查，结合企查查、天眼查等平台确定其是否为同一家企业。输出清洗表格格式参考如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 汽车新产品竞争网络企业名称A | 生产地址A | 汽车产业专利数据企业名称B | 生产地址B | 相似度 | 统一后企业名称 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

1. 对汽车新产品竞争网络中的企业，同时拥有专利的企业，按照第7、8、9、10步确定的原则，计算其3年期、5年期IPCSG和IPCMG网络，并计算同样网络指标。
2. 在“AmIndustryWTOPanel20181128”中新建变量cptoripcmgdgr3yw (三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络degree centrality的均值)、cptoripcmgbtw3yw(三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络shortest-path betweenness centrality的均值)、cptoripcmglcrch3yw均值（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络local reaching centrality的均值）、cptoripcmgcnstr3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络constraint的均值）、cptoripcmgeffsz3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络effective size的均值）、cptoripcmgtrgl3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络effective size的均值）、cptoripcmgclust3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络clustering/clustering的均值）、cptoripcmgntsz3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络network size的均值）、cptoripcmgntdnst3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络network density的均值）、cptoripcmgcliq3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络中最大clique规模的均值）、cptoripcmgeffc3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络中efficiency的均值）、cptoripcmgisolt3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCMG网络中isolate数量的均值）。将相应的计算结果导入面板数据中的相应变量。
3. 在“AmIndustryWTOPanel20181128”中新建变量cptoripcmgdgr5yw (五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络degree centrality的均值)、cptoripcmgbtw5yw(五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络shortest-path betweenness centrality的均值)、cptoripcmglcrch5yw均值（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络local reaching centrality的均值）、cptoripcmgcnstr5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络constraint的均值）、cptoripcmgeffsz5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络effective size的均值）、cptoripcmgtrgl5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络effective size的均值）、cptoripcmgclust5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络clustering/clustering的均值）、cptoripcmgntsz5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络network size的均值）、cptoripcmgntdnst5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络network density的均值）、cptoripcmgcliq5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络最大clique规模的均值）、cptoripcmgeffc5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络efficiency的均值）、cptoripcmgisolt5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCMG网络isolate数量的均值）。将相应的计算结果导入面板数据中的相应变量。
4. 在“AmIndustryWTOPanel20181128”中新建变量cptoripcsgdgr3yw (三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络degree centrality的均值)、cptoripcsgbtw3yw(三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络shortest-path betweenness centrality的均值)、cptoripcsglcrch3yw均值（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络local reaching centrality的均值）、cptoripcsgcnstr3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络constraint的均值）、cptoripcsgeffsz3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络effective size的均值）、cptoripcsgtrgl3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络effective size的均值）、cptoripcsgclust3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络clustering/clustering的均值）、cptoripcsgntsz3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络network size的均值）、cptoripcsgntdnst3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络network density的均值）、cptoripcsgcliq3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络最大clique规模的均值）、cptoripcsgeffc3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络efficiency的均值）、cptoripcsgisolt3yw（三年窗口期汽车新产品竞争网络中某企业竞争对手三年窗口期IPCSG网络isolate数量的均值）。将相应的计算结果导入面板数据中的相应变量。
5. 在“AmIndustryWTOPanel20181128”中新建变量cptoripcsgdgr5yw (五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络degree centrality的均值)、cptoripcsgbtw5yw(五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络shortest-path betweenness centrality的均值)、cptoripcsglcrch5yw均值（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络local reaching centrality的均值）、cptoripcsgcnstr5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络constraint的均值）、cptoripcsgeffsz5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络effective size的均值）、cptoripcsgtrgl5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络effective size的均值）、cptoripcsgclust5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络clustering/clustering的均值）、cptoripcsgntsz5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络network size的均值）、cptoripcsgntdnst5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络network density的均值）、cptoripcsgcliq5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络最大clique规模的均值）、cptoripcsgeffc5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络effciciency的均值）、cptoripcsgisolt5yw（五年窗口期汽车新产品竞争网络中某企业竞争对手五年窗口期IPCSG网络isolate数量的均值）。将相应的计算结果导入面板数据中的相应变量。
6. 在“AmIndustryWTOPanel20181128”中新建变量cptor3ywipcsgcsd表示在三年窗口期汽车新产品竞争网络中某企业的三年期IPCSG向量与它的竞争对手的三年期IPCSG向量之间cosine技术距离的均值。COSINE技术距离计算公式如下：



表示企业i与企业j之间IPCSG向量或者IPCMG向量的COSINE技术距离。其中表示企业i的当期IPCSG矩阵中某个IPCSG分类号k所拥有的专利数量对该企业当期所有专利数量的比值。比如A公司某期IPCSG矩阵如下图所示。则该A对应H04N5/45的为0/2=0，而A对应h04n5/76的为2/2=1。



求出企业i与它的任意一个竞争对手j之间的COSINE技术距离后，再求这些COSINE技术距离的均值，即为i公司cptor3ywipcsgcsd的取值。

1. 在“AmIndustryWTOPanel20181128”中新建变量cptor5ywipcsgcsd表示在五年窗口期汽车新产品竞争网络中某企业的五年期IPCSG向量与它的竞争对手的五年期IPCSG向量之间cosine技术距离的均值
2. “AmIndustryWTOPanel20181128”中新建变量cptor3ywipcmgcsd表示在三年窗口期汽车新产品竞争网络中某企业的三年期IPCMG向量与它的竞争对手的三年期IPCMG向量之间cosine技术距离的均值
3. “AmIndustryWTOPanel20181128”中新建变量cptor5ywipcmgcsd表示在五年窗口期汽车新产品竞争网络中某企业的三年期IPCMG向量与它的竞争对手的三年期IPCMG向量之间cosine技术距离的均值