计算机辅助批量估价在房地产信息系统中的实现

杨振鹏[1] 马亚男[2]

（作者单位：1 清华大学；2 中国房地产估价师与房地产经纪人学会）

摘要：计算机辅助批量估价（CAMA）可用于房地产评税、银行抵押贷款风险分析、在线房价查询等多种用途。本文介绍了一种在房地产信息系统中实现批量估价的方法。该方法以多元回归分析为理论基础，通过数据库二次开发完成评估模型的回归和应用。实验证明，其评估结果符合国际评估师协会（IAAO）对住房批量估价的要求。由于使用通用的数据库查询语言（SQL），该方法具有较好的实践借鉴价值。

关键词：房地产估价；批量评估；计算机辅助；特征价格

一、引言

计算机辅助批量估价（Computer-Assisted Mass Appraisal，简称 CAMA），是借助计算机的数据分析处理能力，按照一定的评估程序高效率地批量执行房地产估价的技术[1]。在国外，CAMA 技术经常用于住宅物业税的税基测算中[2]——计算机根据数据库中存储的可比交易案例资料，通过数据挖掘（Data Mining）手段得到评估模型，并以此评估每一套应纳税住宅的价值。

除了可用于房地产评税以外，CAMA 技术还有很多应用前景[3]。银行可以利用它计算并及时更新每一宗抵押贷款的贷款价值比（LTV），实现风险的动态监控；经纪人可以在客户询问不熟悉的楼盘或区域时，利用 CAMA 系统为客户进行初步估价，从而提高服务的水平；评估机构可以将自动估价的结果作为参考；房地产门户网站可以在 GIS 平台上展现符合用户要求房型的价格，而不再仅仅是楼盘的均价。

本文介绍一种在房地产信息系统中实现计算机辅助批量估价的方法。该方法采用多元回归分析（Multiple Regression Analysis）构建批量评估模型，并结合数据库完成模型的回归和应用。实验证明，其评估结果符合国际评估师协会（IAAO）对住房批量评估的要求。由于使用通用的数据库查询语言（SQL），该方法具有

较好的实践借鉴价值。

二、批量估价所需的基础数据

基础数据对于房地产估价至关重要，数据质量的优劣决定了评估结果的准确性。对计算机辅助批量估价而言，基础数据包括两个组成部分，即住房属性数据和可比案例的交易价格数据。这两个部分相互独立，通过房屋的唯一编码进行关联。

（一）住房属性数据

住房属性数据包括房屋所在的小区、楼栋、单元、楼层、房屋的建筑面积、使用面积、朝向、户型以及楼栋的建筑形式等。收集存量住房属性信息的工作量大，特别是对于没有建立“电子楼盘表”的小区，需进行系统的调查，对需要评估住房的状况建立电子化记录。不过，在个人住房信息系统完善的城市中，可以利用国土、房管、建设等相关部门的信息建立住房属性数据库。

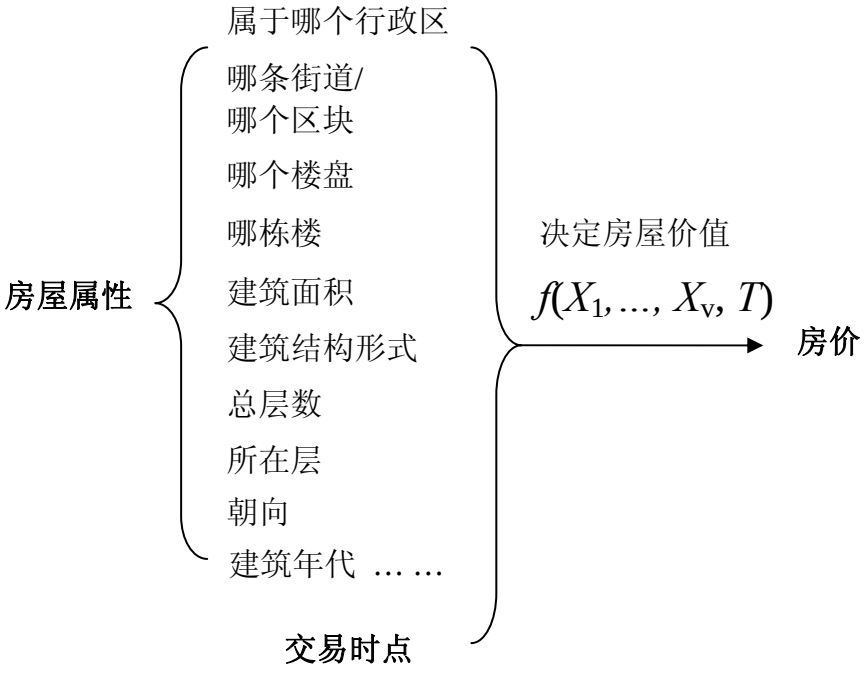


图 1 典型的房屋属性

（二）可比案例的交易价格数据

1. 存量房交易流程

住房交易价格数据的收集则更为复杂，这主要由两方面原因造成。首先，住

房交易过程持续时间较长，在不同的环节会产生不同的价格，因此“交易价格”本身就具有多义性，数据收集渠道也不唯一；第二，房地产交易属于私人市场中的交易，出于多种目的，交易双方可能倾向于隐藏实际价格，导致不同渠道收集的数据相互矛盾，难以判断真实的价格。

具体来说，我国存量住房交易过程可分为搜寻、议价、签订合同、抵押评估、申请贷款、缴纳税费、办理过户、物业交割、放款等若干环节，整个流程如图 2 所示。

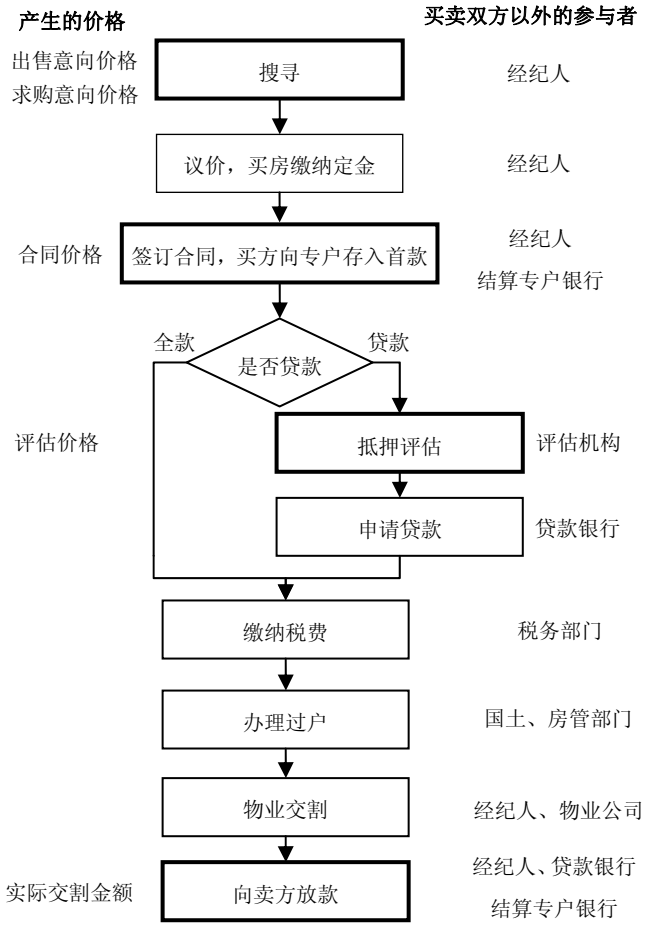


图 2 我国存量住房交易的一般流程

2. 各环节产生的“交易价格”及其特点

可以看到，在交易流程中可能产生多种价格，其中最重要的三者是合同价格、评估价格和实际交割金额。

从理论上讲，存量房买卖合同是双方真实意愿的表达，在买卖双方都充分调查了解房屋状况、市场行情的前提下，合同价格应当是反映房屋在签约时点价值

的最好指标。但是，由于很多交易环节的税费以合同价格作为计税依据，有买房者和卖房者通过订立虚假合同减少税负。这种现象在我国并不罕见，有媒体的调查显示“签阴阳合同已经成为公开的秘密，绝大部分买房人都会以此来避税”[4]。因此，如果采用合同价格作为可比案例价格，必须注意核实价格的真实性。

在抵押评估环节，会产生一个由专业评估机构给出的评估价格。这一价格通常更接近房屋的实际市场价值，但也有报道指出，评估机构同样可能根据买卖双方的要求调整评估结果[5]。

最后，存量房买卖结算专户中的资金往来情况也能在一定程度上反映交易的金额。但这一金额并不必然等于真实的合同额。这是因为双方不一定约定 100% 的交易资金都通过结算专户划转，例如买受人缴纳的定金就有可能直接以现金支付给出卖人，而配套家电和装修款项则有可能并入结算专户划转。

三、计算机辅助批量估价的原理和方法

计算机辅助批量估价可采用多种方法，其中最成熟和流行的当属“多元回归分析法”（Multiple Regression Analysis，简称 MRA）。MRA 是分析因变量如何随着一个或多个自变量的变化而变化的统计方法[6]，它的最简单形式是采用普通最小二乘准则（OLS）的多元线性回归。

（一）采用 OLS 进行回归分析的原理

当采用 OLS 进行回归分析时，隐含的假设是因变量房价 P 和多个自变量 X_1、X_2、…、X_m 呈线性关系，即房价 P 的取值总是与 $X_1 \sim X_m$ 的某个线性表达式 $(b_0 + b_1X_1 + ... + b_mX_m)$ 接近。其中的参数 b_0~b_m 在回归前都是未知的。但是，通过收集足够数量的可比交易案例样本（每个样本都有 X_1~X_m 和 P 的取值），就可以估计出 b_0~b_m 最可能的取值。具体原理如下。

假设有 n 个(X_1, …, X_m, P)样本，样本数量 n 大于自变量的数目 m。对于第 i 个样本，其因变量 P 取值记为 P_i，其自变量取值记为 X_{i1}~X_{im}。“最佳的拟合公式” 中的 b_0~b_m 应当使每一个样本的 P_i 与(b_0+ b_1X_{i1}+…+ b_mX_{im})的差距（残差）尽可能小。而衡量回归公式整体拟合程度的标准，对于 OLS 而言，就是所有样

本残差的平方和，如公式 1 所示：

$$\min.\quad Z(b_0,b_1,\ldots,b_m)=\sum_{i=1}^{n}[P_i-(b_0+b_1X_{i1}+\ldots+b_mX_{im})]^2 \qquad \text{公式-1}$$

可以证明，使样本残差平方和 Z 取最小值时，$(b_0,b_1,\ldots,b_m)$应当为：

$$(\hat{b}_0,\hat{b}_1,\ldots,\hat{b}_m)^{\mathbf{T}}=(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \qquad \text{公式-2}$$

公式 2 采用矩阵向量表示，其中：

$\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^{\mathbf{T}}$　　$\mathbf{X}_i=(1,X_{i1},\ldots,X_{im})^{\mathbf{T}}$

$\mathbf{Y}=(P_1,\ldots,P_n)^{\mathbf{T}}$

利用公式 2，就可通过样本数据 **X** 和 **Y** 计算回归公式中 b_0~b_m 的估计值。

（二）批量估价实践中 OLS 回归的步骤

在批量估价实践中，可以用 OLS 回归分析可比案例交易价格与房屋属性的关系，得到房价的回归公式，从而用于对其他住房价格的评估。

然而，操作中往往会遇到以下障碍：①有一些房屋属性不是数值型的变量（如房屋的朝向）；②交易时点也对交易价格有重要影响；③某些房屋属性对于房价的影响可能不是线性的；④有一些房屋属性对房价没有影响（在统计意义上）。因此，在实践中使用 OLS 时一般要进行以下步骤。

1. 非数值型变量的转换

有些房屋属性的取值不是数值型的，而是从特定列表中选定一个取值。例如房屋朝向，一般是在东、南、西、北、东南、西南、东北、西北、南北、其他等 10 个可能的选项中选择一个。为了采用多元线性回归分析这种属性对住房价格的影响，需要将其转换为一组虚变量。方法是，对每个可能的选项都设置一个虚变量；样本属于哪个选项，它对应的虚变量就取 1，其他虚变量均取 0。

以房屋朝向为例，为了将其数值化，需要在基础数据库中添加 10 个虚变量字段，依次是 D_{E}、D_{W}、D_{S}、D_{N}、D_{SE}、D_{SW}、D_{NE}、D_{NW}、D_{NS} 和 D_{other}。对于一个朝南的房屋，D_{S} 取 1，其他 9 个字段均为 0。

在回归前，统计每个选项出现的频率，出现频率最高的选项对应的虚变量不引入回归模型（避免线性相关），其余虚变量都作为自变量加入回归模型中。

2. 将交易时点转换为自变量

交易时点也需要转化为数值型变量才能引入回归模型，通常采用时间虚变量法。与上面步骤类似，对每个出现交易案例的月份（或季度）设置一个虚变量。样本在哪个月交易，这个月的虚变量就取 1，其他月份的时间虚变量取 0。

3. 因变量和自变量的 Box-Cox 变换

有些因素对于住房价格的影响不是线性的。例如房屋所在楼层，一般情况下中间楼层的房价较高。在这种情况下，需要采用 Box-Cox 变换提高拟合效果。本文采用因变量取对数、数值型自变量加二次项的变换形式。即构造一个房价的自然对数的字段作为因变量；对每个数值型变量，额外添加一个二次项字段（变量的平方），也作为自变量一同加入模型。因此，最终的回归模型类似于公式 3 的形式。

$$\mathrm{Log}(P) = b_0 + b_1X_1 + b_1'X_1^2 + \ldots + b_pX_p + b_p'X_p^2 + c_1D_1 + \ldots + c_qD_q \qquad \text{公式-3}$$

其中 X_1~X_p 是数值型变量，D_1~D_q 是由非数值型属性以及交易时点转化来的虚变量。

4. 自变量的显著性检验

并不是所有房屋属性都和房价有显著的关系。例如小区总面积和住宅价格的关系就可能是模糊的，或者说在统计上关系不显著。加入大量关系不显著的自变量将对回归产生负面影响，因此，需要对每一个自变量进行显著性检验，去除冗余的变量。

显著性检验假设因变量的取值 $\mathbf{Y}$ 是一个以 $\mathbf{Xb}$ 为均值、$\sigma^2\boldsymbol{I}$ 为方差的正态分布随机变量（向量）。根据公式 2，$\mathbf{b}$ 的估计值 $\hat{\mathbf{b}} = (\mathbf{X^TX})^{-1}\mathbf{X^TY}$，是 $\mathbf{Y}$ 的线性函数，因此 $\hat{\mathbf{b}}$ 也是正态分布随机变量，可知其均值和方差-协方差矩阵为：

$$E(\hat{\mathbf{b}}) = E[(\mathbf{X^T X})^{-1}\mathbf{X^T Y}] = (\mathbf{X^T X})^{-1}\mathbf{X^T X b} = \mathbf{b}$$

$$var.(\hat{\mathbf{b}}) = var.[(\mathbf{X^T X})^{-1}\mathbf{X^T Y}] = (\mathbf{X^T X})^{-1}\mathbf{X^T}(\sigma^2 \boldsymbol{I})\ \mathbf{X}\ (\mathbf{X^T X})^{-1} = \sigma^2\ (\mathbf{X^T X})^{-1} \quad \text{公式-4}$$

公式 4 体现了由于样本具有偶然性，**b** 的估计结果 $\hat{\mathbf{b}}$ 也会有随机扰动，这个扰动的标准差被称为标准误差（Standard Error）。如果某个变量系数的标准误差很大，以至于和估计值本身在一个量级上，那么估计值就丧失了意义。这时需要重新考虑是否应当引入这个自变量。本文采用“对数似然比检验”考察所有 t-统计量（估计值与标准误差之比）相伴概率大于 10%的自变量。如果对数似然比的相伴概率亦大于 10%，则从模型中剔除这个变量。

（三）批量估价效果的评价

国际评估师协会（IAAO）在 1999 年的《Standard on Ratio Studies》中确定了几项用于评价批量评估效果的指标，如表 1 所示。

表 1 传统 MRA 模型评估效果

评价指标	计算公式	IAAO 建议范围
单宗案例的评估比率 AR_i	A_i / S_i，其中 A_i 为评估价格 S_i 为实际价格	
简单平均评估比率 *AR1*	$AR1 = \sum_{i=1}^{n} \frac{A_i}{S_i} / n$	0.9-1.1
中位数评估比率 *M*	AR_i 的中位数	0.9-1.1
加权平均评估比率 *AR2*	$AR2 = \frac{\sum_{i=1}^{n}(\frac{A_i}{S_i} \cdot S_i)}{\sum_{i=1}^{n} S_i} = \sum_{i=1}^{n} A_i / \sum_{i=1}^{n} S_i$	0.9-1.1
价格相关差 *PRD*	$PRD = AR1 / AR2$	0.98-1.03
离散系数 *COD*	$\frac{\sum_{i=1}^{n}\lvert AR_i - M\rvert}{n} \div M \times 100$	10 以下

价格相关差 *PRD* 衡量是否存在与价格相关的偏差——如果 *PRD* 低于 0.98，

说明评估模型倾向于高估高价住房、低估低价住房；如果 *PRD* 高于 1.03，说明出现相反的情况。离散系数 *COD* 衡量这一批评估相对误差的分布，*COD* 小于 10 说明相对误差分布在±10%以内。

四、实例

本文采用自行开发的计算机软件 CFMA 实现批量估价。CFMA 可以按照指定方式从数据库中读取基础数据，并自动完成非数值型变量转换、时间虚变量设定、Box-Cox 变换以及自变量的显著性检验，确保通过 OLS 回归得到最佳的房价评估模型。利用回归得到的模型，CFMA 可以对待估案例进行评估，给出评估价格和置信区间。

（一）收集测试数据

本文收集了 573 条二手住房出售信息作为测试数据。地域涉及北京市 4 个特定地区，分别是丰台区的青塔地区、赵公口地区、石景山区的鲁谷地区以及海淀区的玉泉路地区。时间分布在 2011 年的前两个季度。测试数据包括以下字段。

表 2 测试数据的字段设置

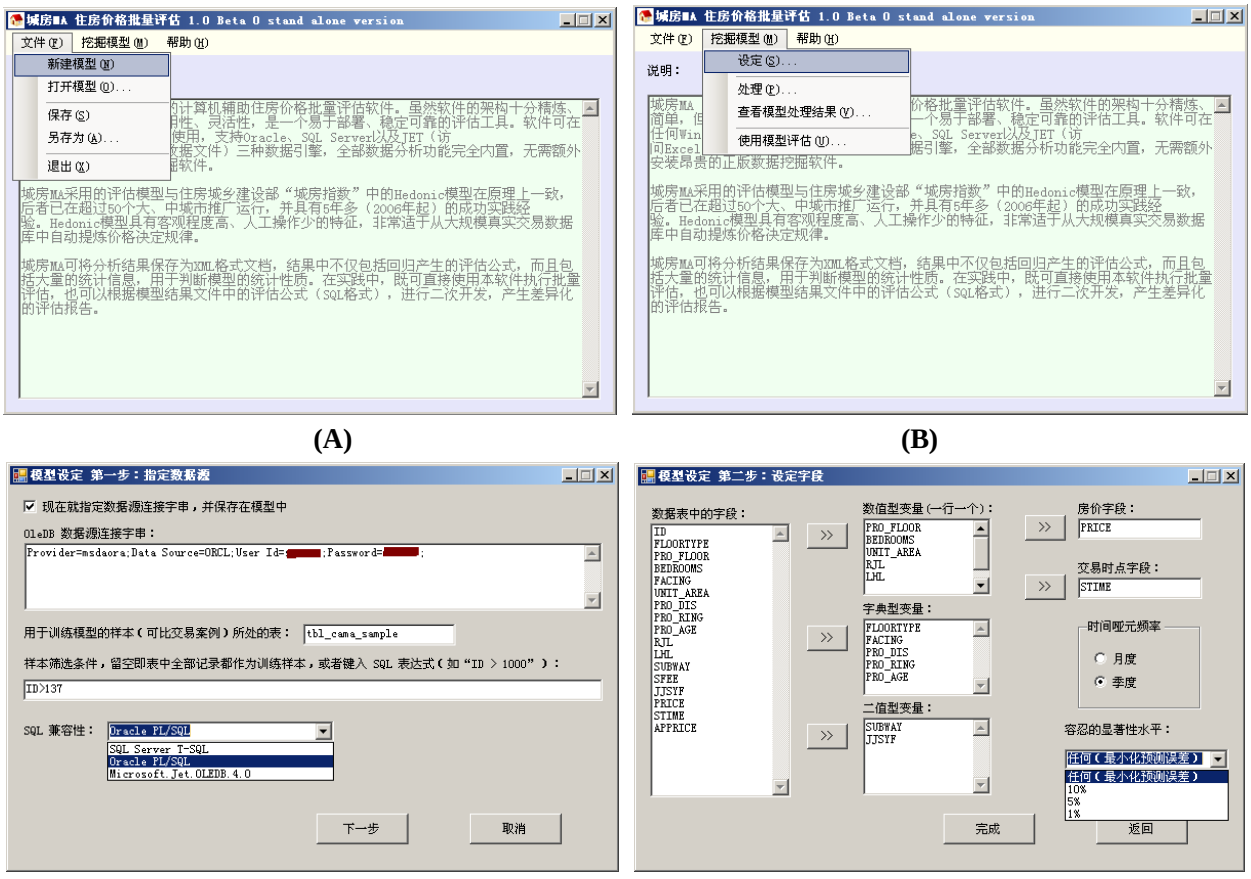
字段名称	含义	取值规则 / 备注
ID	案例编号	流水号
FLOORTPYE	房屋所在楼层	0-低层；1-中间层；2-高层；3-顶层
PRO_FLOOR	建筑总层数	数值
BEDROOMS	卧室数目	数值
FACING	房屋朝向	10-东；20-西；30-南；40-北；13-东南；23-西南；14-东北；24-西北；12-东西；34-南北
UNIT_AREA	建筑面积	单位：平方米
PRO_DIS	所在行政区	1-丰台；2-石景山；3-海淀
PRO_RING	环线位置	2-三环内；3-三环~四环；4-四环外
PRO_AGE	建筑年代	1980；1985；1990；1995；2000；2005；2010
RJL	容积率	单位：100%
LHL	绿化率	单位：100%
SUBWAY	是否靠近地铁	0-没有地铁；1-有地铁
SFEE	物业费	单位：元/平方米/月
JJSYF	是否经济适用房	0-商品房；1-经济适用房
STIME	采样时点	yyyy-mm-dd
PRICE	总价	单位：万元，业主（经纪人）报价
APPRICE	评估价格	用于输出、保存评估结果的字段

本文将上述样本随机分为两个部分，其中 137 条样本作为待评估案例（ID 编为 1~137），另外 436 条样本作为可比交易案例（ID 编为 138~573）。测试数据所在的表名称为“tbl_cama_sample”。

本文使用 Oracle 10g 数据库引擎存储数据，这与目前大多数城市政府的房地产信息系统一致。除 Oracle 以外，CFMA 批量估价软件还支持 SQL Server 数据库引擎和 Jet 引擎。因此部署了 SQL Server 数据库的机构也可以使用 CFMA 进行批量估价；小型企业和个人可以使用 Jet 引擎连接 Excel / Access 数据库。

（二）设定评估模型

打开 CFMA 软件，在文件菜单中选择新建模型，然后在“挖掘模型”菜单中选择“设定”，即可打开模型设定界面，如图 2 所示。



(A) **(B)**

(C) **(D)**

图 2 使用 CFMA 软件设定评估模型的截图

1. 设定要访问的数据库

在设定的第一步中（图 2.C），需要填写访问数据库的连接字串、可比交易案例所处的表、筛选条件以及 SQL 格式。

“连接字串”是访问数据库的命令，它告诉 CFMA 软件应当到哪个数据库中提取数据。具体到某个单位，应当向数据库管理员询问访问本单位数据库的连接字串[①]。

“可比交易案例所处的表”，填写上一步准备好的数据表的名称——“tbl_cama_sample”。

“筛选条件”可用于对表中的记录进行进一步筛选。在本文示例中，使用的可比交易案例是 ID 在 138~573 之间的记录，因此输入条件“ID>137”。

在“SQL 兼容性”下拉框中，根据数据库的类型选择。本文使用 Oracle 数据库，因此选择 PL/SQL；如果使用 SQL Server，应选择 T-SQL；如果使用 Excel / Access，应选择 Microsoft.Jet.OleDB.4.0。

2. 设定评估模型的因变量和自变量

在进入第二步界面时（图 2.D），如果上一步设定的数据库连接正常，会在左侧列表中列出全部字段。在这一步中，需要指明哪些字段是自变量，哪个字段是可比交易案例的价格，哪个字段是交易时点，以及显著性检验的方式。

在本文准备的数据表中，房价字段为 PRICE，价格时点为 STIME，将它们填入相应位置。并选择时间虚变量的频率为季度。

对于取值为数值的属性，如本文中的 PRO_FLOOR、BEDROOMS、UNIT_AREA、RJL、LHL 和 SFEE，填入数值型变量列表中。

对于非数值型字段，取值只有 0/1 两种选择的，如 JJSYF、SUBWAY，填入二值型变量列表；取值有多种可能的，如 FLOORTYPE、FACING、PRO_DIS、PRO_RING、PRO_AGE，填入字典型变量列表。

---

① 如果希望使用 Excel 2003 作为数据库，连接字串应写为“Provider = Microsoft.Jet.OLEDB.4.0; Data Source = "*包含路径的 XLS 文件全名*"; Extended Properties="Excel 8.0;HDR=Yes;IMEX=2";”。但 Jet 引擎在处理大量数据时效率较低，会消耗大量时间。

最后，显著性检验的准则为“最小化预测误差”。

（三）回归计算

设定完成后，可以在“挖掘模型”菜单中选择“处理”，进行回归计算。首先点击“估计计算量”，再点击“开始处理”。计算时间取决于样本量和变量数目的多寡。

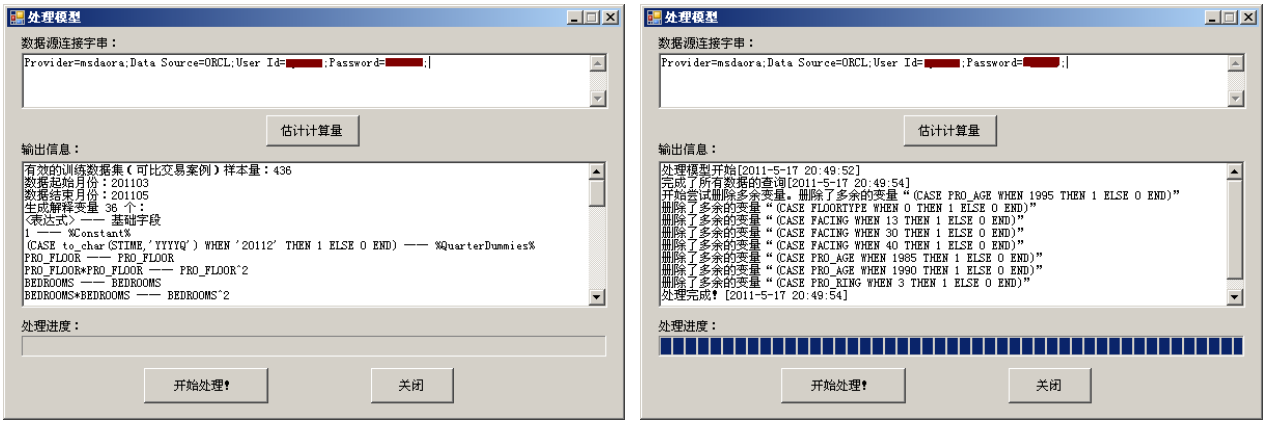


图 3 使用 CFMA 软件进行回归的截图

CFMA 软件根据内置的程序自动完成非数值型变量转换、时间虚变量设定、Box-Cox 变换以及自变量的显著性检验。计算完成后，会在“输出信息”中显示在显著性检验中被删除的变量。

（四）查看评估模型的回归结果

在“挖掘模型”菜单中选择“查看模型处理结果”，可查看回归结果。

图 4 使用 CFMA 软件查看回归结果的截图

在结果中，可显示各个变量的系数估计值、标准误差、t-统计量，以及反映模型整体效果的统计量，如 R-平方、F-统计量、对数似然值、AIC、SC 等。其中，“1 倍标准误差范围”给出了使用本模型进行评估所具有的误差。在示例中，误差为 -8.90% ~ + 9.77%，符合 IAAO 确定的误差在±10%之内的标准。

（五）利用模型进行批量估价

在“挖掘模型”菜单中选择“使用模型评估”，输入待估案例所处的表、存放评估结果的字段、筛选条件，即可执行批量估价。

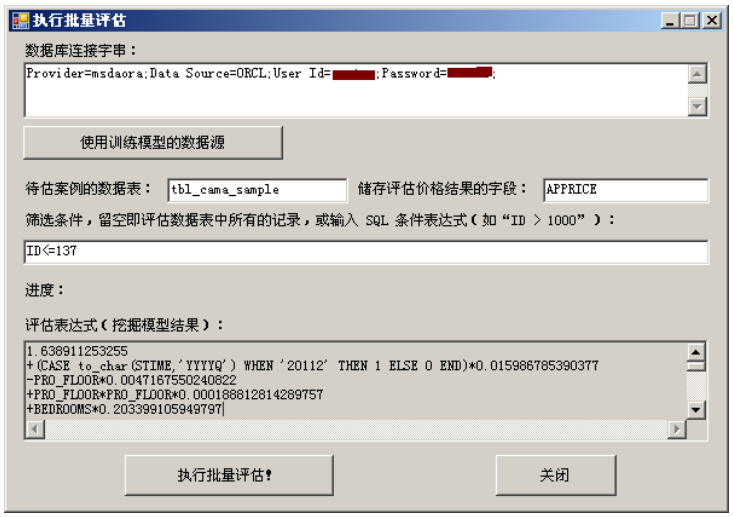


图 4 使用 CFMA 软件查看回归结果的截图

在本文示例中，以 ID 为 1~137 的记录为待估案例，因此在筛选条件中输入“ID<=137”。将评估结果存放在“APPRICE”字段中。完成后，评估结果会存储在数据库相应字段中，如图 5 所示。

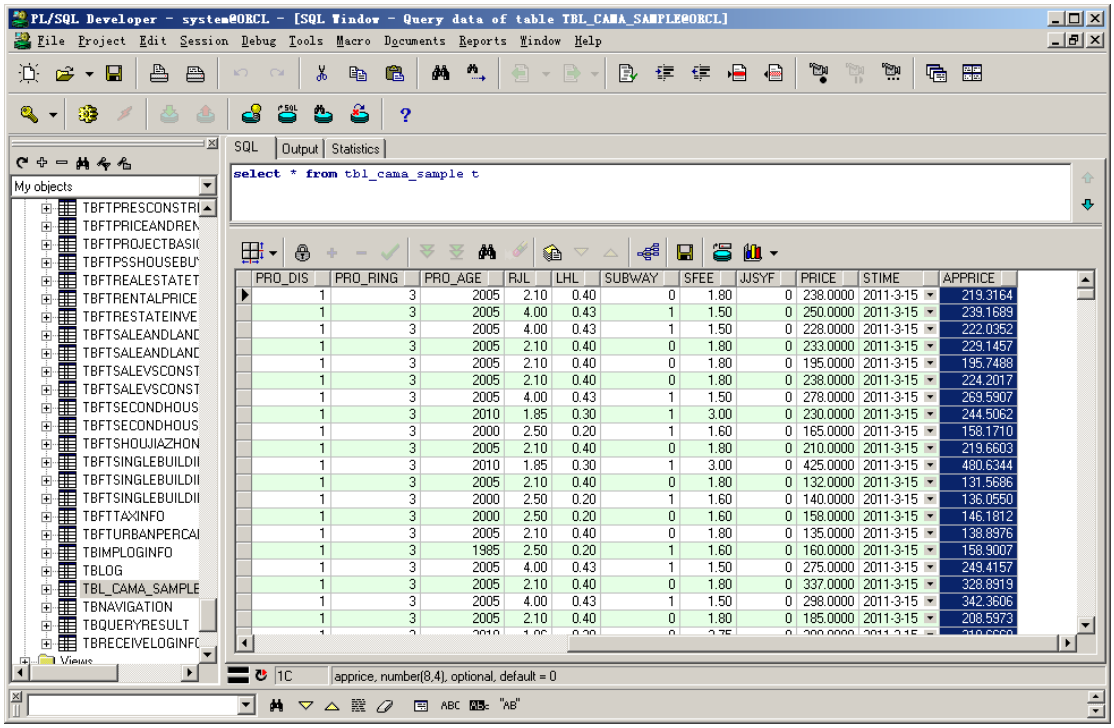


图 5 在数据库中查看批量估价结果

（六）根据 IAAO 的标准衡量评估效果

依据表 1 中的公式，可以评价本次批量估价的效果，如表 3 所示。可以看到，本次批量估价效果非常理想。

表 3 评估效果

	AR1	AR2	M	PRD	COD
本次批量估价	0.984	0.977	0.981	1.007	5.62
IAAO 的要求	0.9-1.1	0.9-1.1	0.9-1.1	0.98-1.03	< 10
是否达标	√	√	√	√	√

五、结语

本文介绍了一种在房地产信息系统中实现计算机辅助批量估价的方法。该方法采用多元回归分析（Multiple Regression Analysis）构建批量评估模型，并通过自行开发的批量估价软件 CFMA 完成模型的回归和应用。实验证明，其评估结果符合国际评估师协会（IAAO）对住房批量评估的要求。由于使用通用的数据

库查询语言（SQL），该方法具有较好的实践借鉴价值。

参考文献

[1] 任作风, 廖俊平. 计算机辅助批量评估法（CAMA）在物业税估价中的应用[J]. 中国房地产估价师, 2005(1), 75-77.

[2] 纪益成, 王诚军, 傅传锐. 国外 AVM 技术在批量评估中的应用[J]. 中国资产评估, 2006(3), 13-18.

[3] 刘洪玉. 计算机辅助批量评估：国际经验与我国的应用前景[J]. 中国房地产估价与经纪, 2007(6), 20-22.

[4] 李可,王逸吟. 阴阳合同：省钱背后有风险[N]. 光明日报,2010-11-11(009).

[5] 赖伟行. 二手房身价评估有玄机[N]. 广州日报,2009-03-29(005).

[6] John D. Benjamin, Randall S. Guttery and C. F. Sirmans. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation[J]. Journal of Real Estate Practice and Education, 2004(7), pp. 65-77.