基于大数据分析的橡胶期货交易策略研究[1]

参赛队员：岳艳涛　章雅婷　张宇

指导老师：　　邓晓衡　　

参赛单位：　　中南大学　　

提交日期：　二零一五年六月　

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类本科生组三等奖。

摘要

商品期货是买卖双方在未来某个约定的日期按约定价格进行交易的标准化协议。期货交易所每秒钟提供两笔交易品种的实时数据，如何对海量数据进行分析并对期货价格做出预测，以期获取稳定的收益成为商业与学术界关注的重点研究问题。本文（1）通过统计聚类等挖掘方法对期货数据进行分析，并对挖掘的指标合理性进行检验；(2）提出周期内的持仓时长概率分布，构造时间与收益的权变函数；(3）利用多种方法对价格波动周期分析，分别得出期货交易价格的短中长三种周期。(4）充分挖掘数据信息，建立综合价格预测与交易模型，并通过实验验证了模型具有很好的效果。

首先，根据现有交易数据（成交量、持仓量、总量、买一价、买一量、卖一价、卖一量）确定影响价格的因素。通过成交价与 7 个因素的散点图和相关系数分析，发现成交价与 B1 价 S1 价相关性较高，然后利用 ADF 单位根和 Johansen 检验法对数据原序列、一阶差分序列进行平稳性和协整检验，采用 Granger 因果关系检验确定具有滞后性的影响因素。在对价格的波动方式分类时，首先对折线图分析，利用低通滤波器消除高频随机波动，得到成交价和持仓量波动短周期分别为 9.5、11.62min，而成交价中周期为 4.51h。最后，通过周期图法与傅里叶级数法，计算S_τ^2在$\tau=2$最大，即长周期为 2 天。根据成交价、成交量、持仓量将波动方式分为 8 类，并对市场交易行情分析。同时创新性提出周期内的持仓时长概率分布，构造时间与收益的权变函数。

本文将期货价格的预测分为两部分。首先对橡胶价格进行短期预测，建立自回归与分布滞后模型，用 Eviews6.0 和 SAS9.1 软件进行求解和相关性检验。同时利用小波神经网络，对历史数据进行挖掘和模拟分析，进行 200 和 500 次训练，预测结果的相关系数达到 0.9493，比较预测值与真实值可知模型对短期预测精度高。然后对波动价格长期预测建立阻尼衰减趋势指数平滑模型，并对价格进行预测与误差分析。

最后，建立投资收益模型，以期获得最优收益。从期货交易买卖角度对问题进行分析建立模型，利用 0-1 规划作为约束，以收益$TR_{\max}$作为优化目标，通过价格波动预测的结果和实时数据，建立动态规划模型，利用 3 进制编码（代表每个交易点未交易、买、卖）遗传算法，求解预测价格下的 30 个交易点的交易策略，以确定每个交易点开空单还是平仓。结论表明，基于期货交易数据的短期分析预测模型与交易策略具有较好的精度和稳定性。

关键字：平稳性检验、低通滤波器、波动周期、价格预测、投资收益模型、3 进制编码遗传算法

一、问题的提出

商品期货交易所能够提供正在交易品种的实时交易数据，每秒钟为二笔相关数据。如何利用大数据分析方法进行价格预测，市场判断，从商品期货的交易中获取稳定的收益对于交易者越来越重要。附件中数据文件包含 2012 年 9 月 19 个交易日的橡胶交易数据。以所提供数据为基础，分析下列问题：

1、 通过大数据分析，寻找价格波动的相关指标，对价格的波动分类。

2、 交易者往往是根据交易所提供的实时数据，对价格的后期走势做出预测来决定是买入还是卖出。建立合理的橡胶价格波动预测模型；

3、 利用相关结论，寻找最优交易策略。

二、研究现状背景及存在的问题

2.1 期货和橡胶期货的交易方式

期货是现在进行买卖，但是在将来进行交收或交割的标的物。天然橡胶价格影响因素多，价格波动频繁而且剧烈。所以，天然橡胶生产商、流通商和消费商都迫切需要通过期货市场进行套期保值，以达到转移风险的目的，与此同时，许多投资者也希望在频繁的价格波动中获取投机收益。

期货市场实行 T+0 交易制度，即当天买入的期货合约在当天就可以卖出。投资者不但能在一天能进行多轮的买卖开平仓，增加资金的周转率，而且在行情发生较大波动时投资者还可以快速改变头寸方向，避免风险从中获利。

2.2 中国期货和橡胶期货的交易发展

我国期货市场已历经 20 多年的发展历程，商品期货交易的品种迅速增加，吸引了大量交易者的参与。

可见中国期货期货市场整体走势不平稳,经历了大起大落、波澜起伏的过山车行情。

由此可见,中国期货市场整体的交易状况呈现出较为明显的阶段性特征,如何对海量数据进行分析,以期从商品期货的交易中获取相对稳定的收益成为交易者非常关注的问题。

2.2 橡胶期货的预测与交易研究现状

付建岭[1]建立了基于反馈灰色Markov理论的期货价格预测模型，但所用的Markov模型和GM(1，1)模型有较多缺点。刘轶芳[2]的基于期货价格预测模型虽然在金融领域有着的广泛应用，但在我国期货市场中的应用很少。向小东[4]采用小波神经网络方法对石油期货价格进行了预测。但在研究中没有对原始的期货价格数据进行处理,实际上由于期货价格数据反映了每天的随机事件和长时期的趋势，需要对原始数据进行预处理，以提高预测的精度。

同时，已有的研究中，弱化甚至忽略了交易中存在的波动周期的分析，没有对投资者心理因素所造成的的买卖时长的概率分布做出必要的分析。

三、模型构建的假设与前期准备

3.1 数据的来源

根据中国期货官网发布的 2012 年 9 月份的橡胶期货价格的研究报告数据和指导老师橡胶期货交易所得到的相关数据。本文就每天的 200000 组交易数据，包括日期、时间、成交量、持仓增减、B1 价、B1 量、S1 价、S1 量的信息，选择数据挖掘的相应的算法和预测的方法，构建模型。

3.2 问题分析方法梗概

商品期货交易所属于典型的海量数据的统计问题，对海量数据通过正确的方法进行挖掘，深刻理解数据与期货交易的关系，寻求最优的投资策略，获取有效价值，提高预测精度并进行收益与风险控制是本文的研究重点。

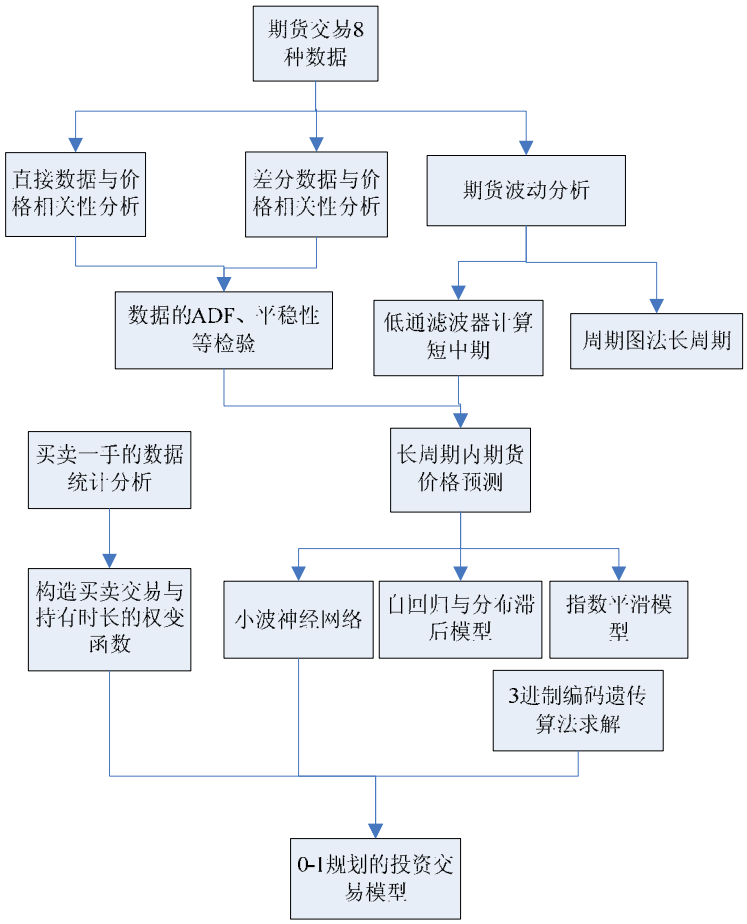


图 1 问题分析总流程

3.3 若干假设

橡胶具有一定的短周期性波动特征,橡胶供求因素、货币政策、汇率政策是长期影响期货市场价格的根本性因素,而自然环境因素、政治因素则是引起天然橡胶期货市场价格短期波动的重要因素。本文提出下列假设：

1.不考虑政府政策，政治、自然、金融货币因素等对期货价格变动造成的影响。

2.交易者能够根据最大收益原则进行投资。

3.该橡胶期货交易数据真实可靠并且具有代表性。
4.期货交易的价格波动存在着不同的波动周期与波动形式

3.4 符号的约定与说明

表1：符号与说明

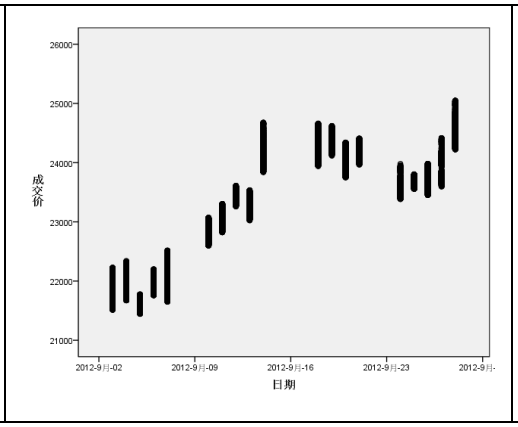
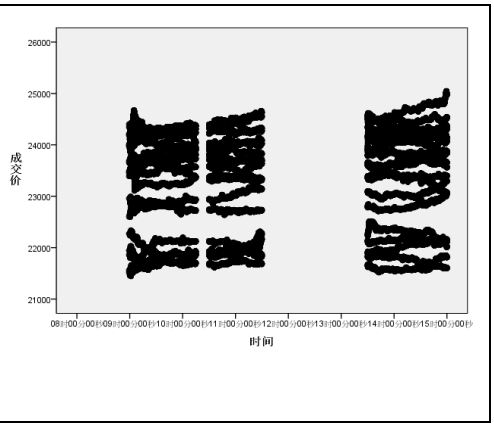
p_t	成交总量	NG_t	交易总量
$\hat{\sigma}_t$	单位时间价格波动估	NC_t	持仓总量
F_i	投资金额	$P_{t,h}$	单位时间成交价极大值
C_i	卖出量	$P_{t,l}$	单位时间成交价极小值
B_i	买入量	R_i	收益率
NBL	买一量	TR	收益总和
NS	卖一价	NB	买一价
NSL	卖一量	NC	持仓量
Ω	周期	NL	成交量
NJ	成交价		

本文使用参数较多，其他公式和符号在具体模型中再做说明。

3.5 数据的预处理

论文[1][2][2]研究了橡胶期货价格和交易量、持仓量的相关性，论文[4][5][6][7]多角度对天然橡胶期货与现货价格关系研究。本文利用 SPSS 软件绘制出了成交价与日期、时间、成交量、持仓增减、B1 价、B1 量、S1 价、S1 量八个因素的散点图并计算相关系数矩阵。如表 2 所示。

表2：成交价与八个因素的散点图

日期		时间	

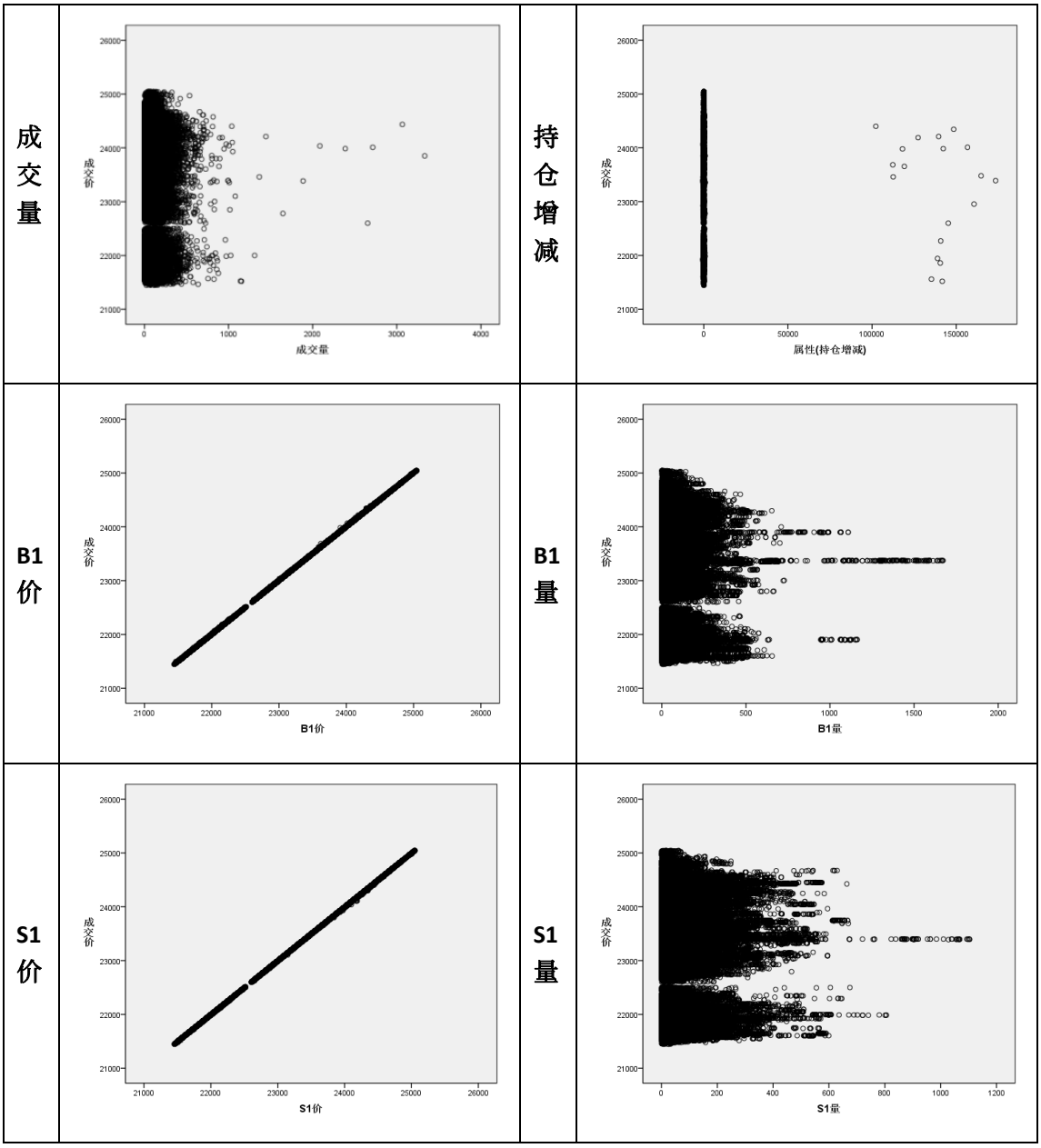


表 3：相关系数表

	日期	时间	成交价	成交量	总量	B1价	持仓增减	B1量	S1价	S1量
日期	1.000									
时间	.006	1.000								
成交价	**.841**	.049	1.000							
成交量	-.025	-.040	.001	1.000	-					
总量	-.092	.896	.046	-.039	1.000					
B1价	.841	.049	**1.000**	.001	.046	1.000				
持仓增减	-.001	-.011	-.001	.165	-.014	-.001	1.000			
B1量	-.040	.050	-.040	.038	.046	-.040	-.001	1.000		
S1价	.841	.049	**1.000**	.001	.046	1.000	-.001	-.040	1.000	
S1量	-.046	.058	-.033	.036	.056	-.033	-.003	-.029	-.033	1.000

由散点图和相关系数表可发现，成交价与B1价 S1价有很高相关性，成交价与日期也有显著关系。同时，许多变量之间直接的相关性比较强，即可能存在信息上的重叠。

四、基于数据分析的预测与交易模型

4.1 期货价格波动的因素分析

4.1.1 影响因素的分析与数据处理

由数据知总量 P_t 与成交量 NL 有：$P_t = P_{t-1} + NL$，成交价由 S1 价和 B1 价决定，S1 量与 B1 量有直接关系，数据 BS 与成交价的变化有关系，因此考虑到其相关性，排除这六个变量的影响。T+0 交易规则，使得期货交易价格变化比较快，应考虑各种因素相关的指标对期货价格的影响具有滞后性。对于变量滞后项能否引入采用 Granger 因果关系检验。

4.1.2 因素选取的平稳性分析与检验

本部分选取成交价，成交量，持仓量作为变量，以一分钟为单位时间的数据进行分析，取一分钟内成交价的平均值，一分钟内成交量的总和，即时的持仓量作为处理后的数据。

多时间序列变量的分析方法要求时间序列是平稳的，需要对相对变动序列的平稳性进行检验。运用 ADF 单位根检验法对序列 NJ,NL,NC 的平稳性进行检验，Eviews 检验结果如下：

表4：原序列检验结果

变量	ADF 值	5%临界值	P 值	是否平稳
成交价（NJ）	1.799953	-1.942688	0.9827	不平稳
成交量(NL)	-3.212833	-1.942677	0.0014	平稳
持仓量(NC)	-0.552243	-1.942688	0.4767	不平稳

ADF 检验值小于显著水平为 5%时的临界值，就可以认为该时间序列不存在单方根，即时间序列是平稳的。由表可知 NJ，NC 序列不平稳。对序列进行一阶差分，结果见表 5：

表5：一阶差分序列检验

变量	ADF 值	5%临界值	P 值	是否平稳
DNJ	-11.20823	-2.878212	0.0000	平稳
DNL	-15.22569	-2.878212	0.0000	平稳
DNC	-3.947060	-2.878212	0.0022	平稳

从一阶差分后的 ADF 检验结果来看，三个检验统计量均小于显著水平 5%，即认为成交价，成交量，持仓量的一阶差分序列都不存在均方根，是平稳的时间序列。协整检验中有三个变量，使用 Johansen 检验法。对序列协整检验，结果如 6：

表6：协方差检验

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.237404	55.96503	29.79707	0.0000
At most 1	0.056997	10.16155	15.49471	0.2685
At most 2	0.001440	0.243585	3.841466	0.6216

注：“*”号表示在5%的显著水平下拒绝原假设，说明三个变量之间存在长期的稳定的协整关系。

根据上文成交价，成交量，持仓量的一阶差分序列都是平稳的时间序列，然后运用 Eviews 对三个变量进行 Granger 因果关系检验，Granger 因果关系检验的滞后系数确定取决于 AIC 准则检验部分结果如表 7

表7：Granger因果关系检验

Null Hypothesis:	Obs	F-Statistic	Prob.
NL does not Granger Cause NJ	171	5.26728	0.0017
NJ does not Granger Cause NL		1.18814	0.3160
NC does not Granger Cause NJ	171	0.91387	0.4356
NJ does not Granger Cause NC		1.61292	0.1884

由表中的概率可知，原假设 NL 变化不是 LJ 变化的格兰杰原因成立的概率为0.0017，认为 NL 变化是 LJ 变化的格兰杰原因。由此类推，价格波动与成交量有关。

4.2 周期图法期货波动周期确定

在对橡胶期货价格，成交量，持仓增减波动分析中，得出了一些隐藏在数字背后的规律，据此，有如下的分析方式。

首先，对橡胶期货各指标五分钟的走势进行分析。整理出 9 月份成交价、持仓量，成交量折线图，结果如图 2 所示。

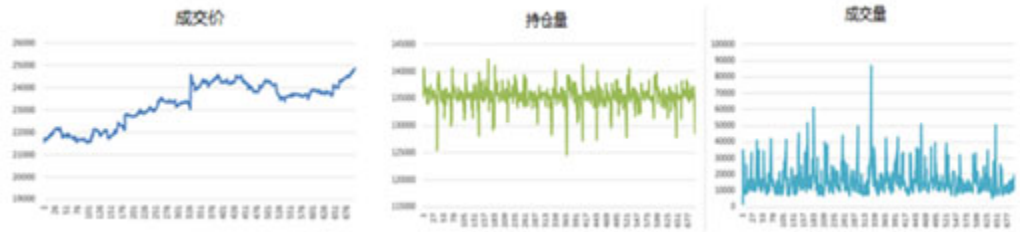


图 2　9 月份每五分钟的成交价、持仓量、成交量

4.2.1 价格滤波的峰谷值法短周期确定

由图 2 可知在 9 月份期货持仓量的增减都有一定的波动和周期性。论文[10][11]中对杂乱无章的经济数据，滤去了随机扰动部分。同理本文也滤除高频扰动，得到价格低频周期波动。

常用的滤波算子[12]有移动平均算子、High-Pass滤波算子、Low-Pass滤波算子等。低通滤波只允许移动很慢的低频 $([-w^*, w^*])$ 信息通过的滤波算子。对称性要求为：$U[w]=U[-w]$，由频反函数，构造滤波算子：

$b_h=\frac{1}{2c}\int_{-c}^{c}U(k)e^{jkh}dk$ 可通过逆傅立叶变换求得：

$b_h = \frac{1}{2c}\int_{-c}^{c} U(k)e^{jkh}dk = \frac{1}{2c}\int_{-k}^{k} e^{jkh}dk$ 即 $b_0 = k/c$，$b_h = \sin(kh)/kh$ 定义

$b(L) = \sum_{h=-\infty}^{\infty} b_h e^{jkh}$ 。

本文对成交价进行 Low-Pass 滤波分析，去除价格波动中随机扰动的部分，得到长期趋势，用 Matlab 处理前后的图形：

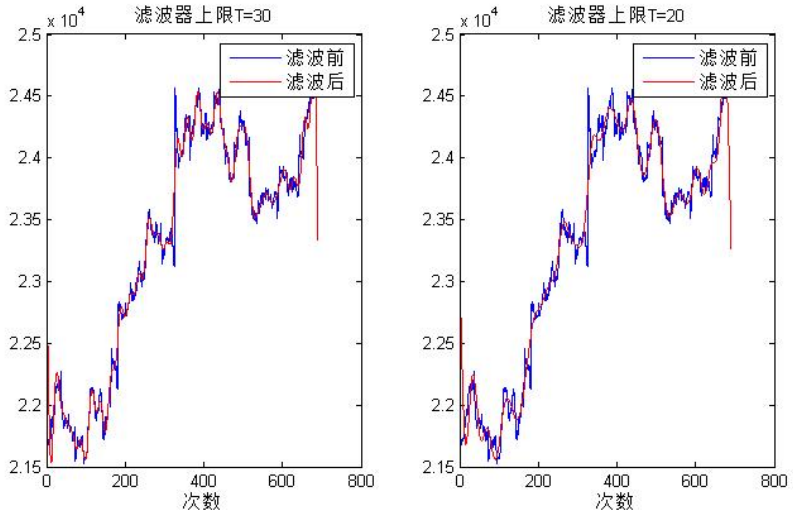


图 3 期货价格滤波处理

由图可知，滤波后数据的趋势得到了很好的体现。接着分析数据总体的波动特征，这里采用简单的经典分析。

表 8：成交价周期转折点

峰值	34	97	241	324	361	408	473	533	581	652
谷值	20	56	118	278	336	383	445	509	552	602
上涨期	41	41	123	46	25	25	28	24	29	50
下跌期	0	22	21	37	12	22	37	36	19	21
周期	0	36	62	160	68	47	63	64	43	50

由数据可知，最后一个峰值不具有参考价值。成交价的周期除了一个数据异常大外，其他数据均近似，异常大点为数据上升较快的区域，为了避免上升区域对数据波动周期分析的影响，本文以所测周期去最大值后的平均值作为周期进行分类。单位时间为 5 分钟，周期为

$$\Omega_1 = (62+68+47+63+64+43+50+36)\div 8)*5\div 60 \approx 4.51\text{小时}$$

为寻求期货价格的短时间波动，本文选用 9 月 3 号上午每秒两笔的数据进行分析，作出其成交价，成交量，持仓总量折线图

图 4 成交价，成交量，持仓总量折线图

从期货成交价的走势来看总体呈缓慢增长的趋势，并具有波动性，同时期货持仓量也有一定的波动和周期性，而成交量并没有较为明显的长期趋势。

同理，为寻求期货价格短时间的波动方式及周期，本文对成交价，持仓量指标进行滤波分析。用 Matlab 处理数据后得到滤波前后的图形。

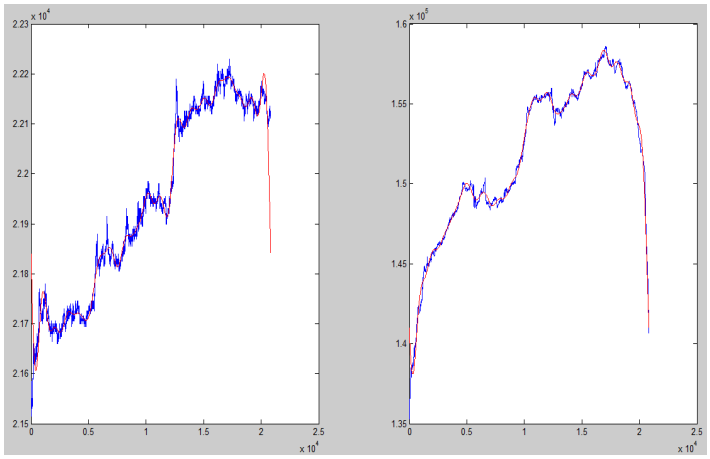


图 5 成交价，持仓量滤波前后图形

将处理后的期货成交价与持仓量进行标准化处理后的图形趋势对比

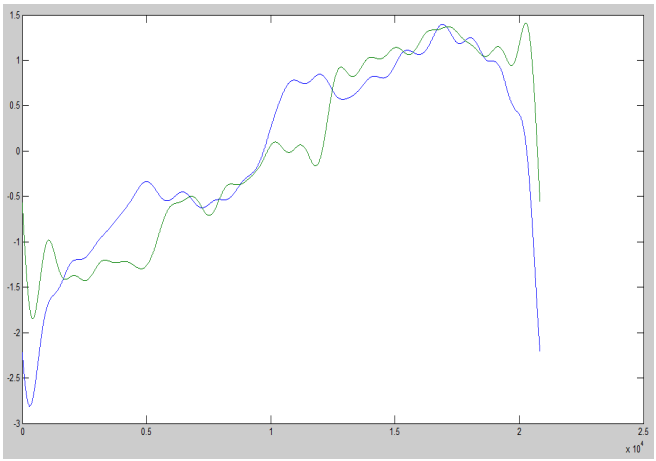


图 6 滤波后的成交价和持仓量

可看出两者的总体趋势近似，期货持仓量对成交价的波动影响在某些区域有明显的滞后性。由图可知图形的拟合程度很高，接着对两指标数据波动进行分析。

表 8 期货成交价周期转折点

峰值	1065 2094 3331 4124 6825 8412 10192 11206 12832 14062 15063 17153 19157 20264
谷值	420 1736 2533 3768 4794 7523 8667 10739 11807 13301 14385 15576 18649 19695

上涨期	645 358798 356 2031 889 1525 467 1025 761 678 1577 508 569
下跌期	0 671 439 437 670 698 255 547 601 469 323 513 1496 538
周期	1316 797 1235 1026 2729 1144 2072 1068 1494 1084 1191 3037 1048

表 9 期货持仓量周期转折点

峰值	5005 6464 7925 10949 11980 14180 15508 18044 18950
谷值	302 5836 7243 8131 11377 12910 14531 15984 17613 18843
上涨期	4703 628 682 2818 603 1270977 938 431 107
下跌期	0 831 779 206 428 930 351 476 691 799
周期	5534 1407 888 3246 1533 1621 1453 1629 1230 0

同理，去除一些异常大的周期数据后得到，期货成交价周期为

$$\Omega_2 = (1316+797+1235+1026+1144+1068+1494+1084+1191+1046)\div 10 \div 2 \div 60 = 9.5\text{分钟}$$

期货持仓量周期为

$$\Omega_3 = (1407+888+1533+1621+1453+1629+1230)\div 7 \div 2 \div 60 = 11.62\text{分钟}$$

所以最终对橡胶期货价格的波动方式进行简单的分类如下：橡胶期货的长期波动周期为 4.51 小时；短时间波动以成交价周期为 9.5 分钟，以期货持仓量来看为 11.62 分钟。在一个周期内，期货持仓量波动周期对期货成交价周期影响有一定的滞后性。

4.2.2Fourier 周期图法隐含长周期确定

周期图法是用试验周期配合实际序列确定隐含周期的方法。论文[13][22]分别对股市波动进行了分析收到较好效果，设对有相等时间间隔的周期函数：$Z_t = C_i Sin(\frac{2\pi t}{T_i} + \varphi_i)(i = 1,2,\cdots\cdots,n)$ 将其展开成傅立叶级数，其一阶谐波的傅氏系数为：

$$A_\tau = \frac{2}{K\tau}\sum_{t=1}^{K\tau} Z_t \cos\frac{2\pi t}{\tau},$$

$$B_\tau = \frac{2}{K\tau}\sum_{t=1}^{K\tau} Z_t \sin\frac{2\pi t}{\tau}$$

式中 Z_t 为每时刻的成交价，K为序列长度n所包含的最大整倍数，相位和振幅为：

$$\varphi_\tau = \operatorname{arctg}\frac{A\tau}{B\tau},$$

$$C_\tau = \sqrt{A^2\tau + B^2\tau}$$

其中 τ 为试验周期。为了确定序列的真正周期，令 $S_\tau{}^2 = A_\tau^2 + B_\tau^2$，取不同的 τ 进行试验，并使 S_τ^2 达到极大值时的 $\tau = T$，平均成交价 H 及其方差 σ^2 的计算：

$$H=\frac{\sum_{i=1}^{n}Y_i}{n},\sigma^2=\frac{\sum_{i=1}^{n}(Y_i-H)^2}{n}$$

平均成交价 H=23363.68， $\sigma^2=955281.2$

由于期货交易市场的开放以天为最小单位，则取每日最终的成交价格进行周期分析。且因每周开放5天，所以在此认为价格波动周期介于一天与五天之间。

即:计算 S_τ^2 $\tau=1,2,\cdots,5$ 如表二。

表10 隐周期相关参数计算表

τ	1	2	3	4	5
K	19	8	6	4	3
A_τ	-9.57363×10^{-13}	186.875	22.77778	15.625	35.03199
B_τ	-1.7793×10^{-12}	-5.31292×10^{-14}	-15.8771	-102	42.89998
S_τ^2	4.08246×10^{-24}	34922.26563	770.9105	10648.14	3067.649

由表二可见 $\tau=2$ 时， S_2^2 达最大值，即第一隐含周期为2天。

4.3 期货的波动分类与持有时间偏好分析

4.3.1 聚类分析法的波动分类

在对价格波动方式进行分类时，主要考虑成交量、持仓量和价格走势之间的关系。绘制成交价随时间变化的折线图、成交量随时间变化的折线图、持仓量随时间变化的折线图。

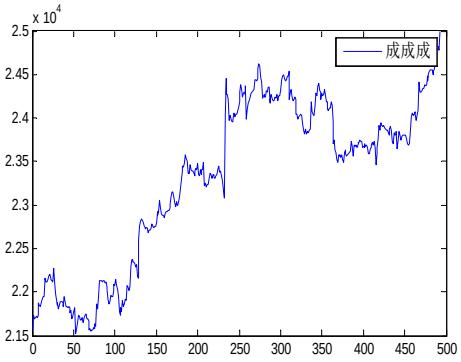


图 7 成交价随时间变化的折线图

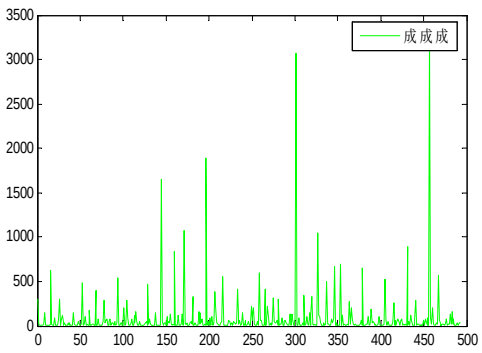
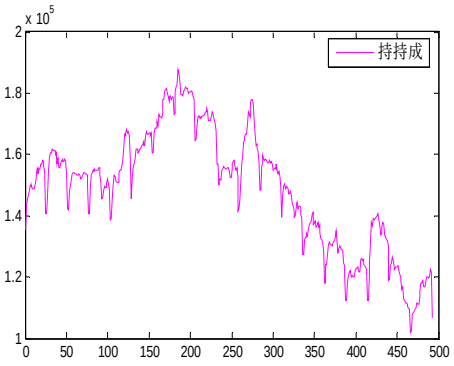


图 8 成交量随时间变化的折线图



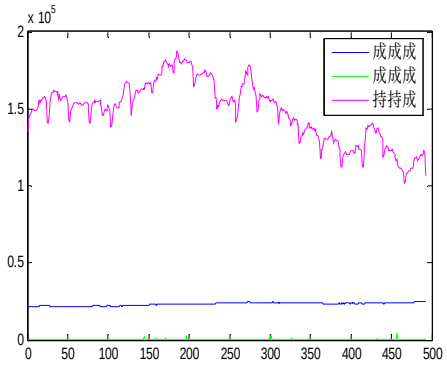


图 9 持仓量随时间变化图　　图 10 成交价、量与持仓量随时间变化图

为了对指标进行分类，采用了聚类方法，选择 R 型聚类分析，即选择成交价、日期、时间等 10 个变量进行聚类分析，并选聚类方法组内链接，以 Pearson Correlation 相关作为测量区间的相关性方法作出相关性矩阵和聚类图。

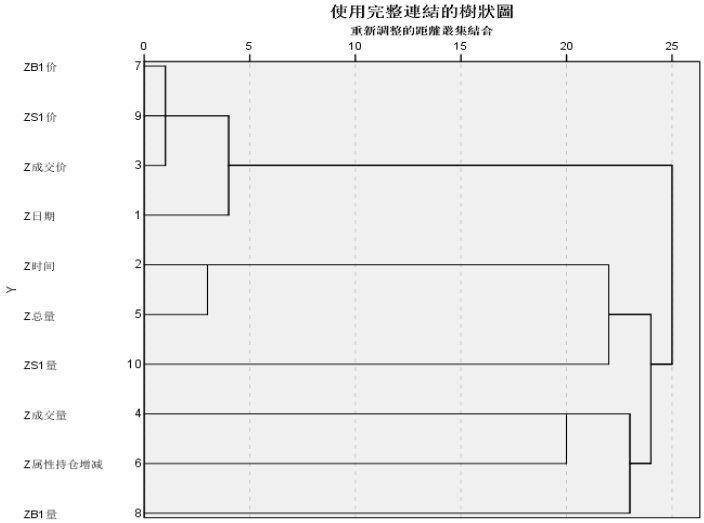


图11 聚类分析图

共分为8类，如下：

表11 各从集組員

观察值	8 从集	7 从集	6 从集	5 从集	4 从集	3 从集	2 从集
日期	1	1	1	1	1	1	1
时间	2	2	2	2	2	2	2
成交价	3	3	1	1	1	1	1
成交量	4	4	3	3	3	3	2
总量	5	2	2	2	2	2	2
持仓增减	6	5	4	3	3	3	2
B1价	3	3	1	1	1	1	1
B1量	7	6	5	4	4	3	2
S1价	3	3	1	1	1	1	1
S1量	8	7	6	5	2	2	2

通过对图 7、8、9、10 的观察和分析可知，价格的波动方式可分为 8 类。

表 12 价格波动方式的市场分析

价格	交易量	持仓量	市场趋向
上涨	增加	上升	坚挺：新开仓增加，多头占优
上涨	减少	上升	瘦弱：新开仓增加，空头占优
下跌	增加	下降	瘦弱：平仓增加，空头迈入平仓占优
下跌	减少	下降	坚挺：平仓增加，多头卖出平仓占优
上涨	不活跃	上升	坚挺：多头占优的情况下平仓减小
上涨	增加	上升	瘦弱：空头占优的情况下平仓减小
下跌	不活跃	下降	空头被逼平仓——空头可能在高位回补
下跌	增加	下降	多头被逼平仓——多头可能在低位回补

4.3.2 期货买卖时间长度的研究

由于期货的交易和收益存在一定的时间关系，投资者可以选择短期持仓后卖出，也可以选择长持仓时间后卖出。此现象解释为投资者对长持仓时间的高收益、短持仓时间的低收益和时间代价三者的综合考虑。

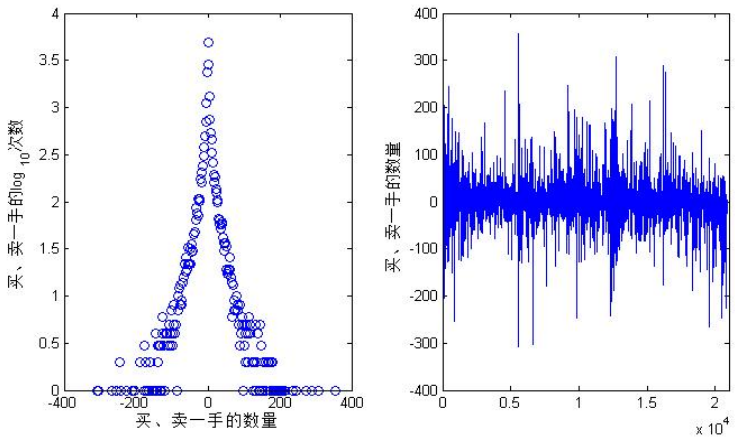


图 12 买卖分布图

根据数据推断，在一个周期内期货交易者可能有不同的持仓时间偏好，对一个周期内的时间长度分割为 6 个时间交易点。即持仓时间相对于研究周期内具有，很短、较短、短、长、较长、很长六种方式。

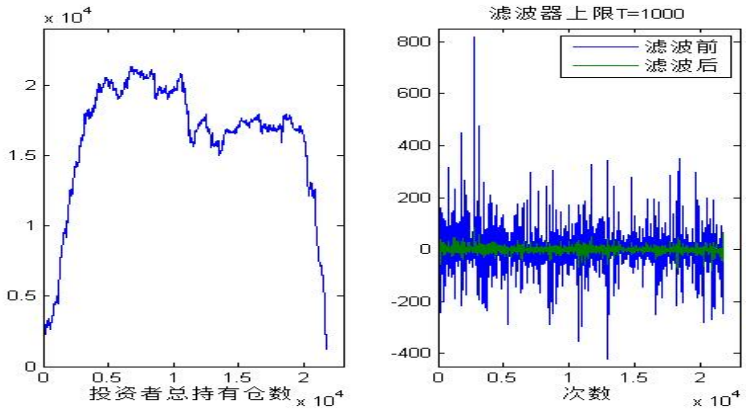


图13 总买卖量变化与周期图

将期货交易状态分为两类。第一类 6 个交易点在交易开始后，买卖一手量会有明显变化（或增或减），但无论该交易周期买卖数量的变化规律如何，该周期结束时净持仓量保持不变或在平衡线上下轻微波动。

考虑到以上情况，构造 S 形增长曲线：

$$f(x)=\begin{cases} a\sqrt[3]{x-\beta}+\gamma, x\geq 0 \\ 0 \qquad\qquad\quad ,x<0 \end{cases}$$

其中 a,β,γ 为待定的常数,当 $x=1$，即假设期货的购入与卖出时间 T 与价格周期相等。

假设量化为 6 个交易点的交易量，根据滤波结果的大量买卖统计，很短、较短、短、长、较长、很长的对应概率值分别为 0.05，0.22，0.41，0.602 0.748，1，即在较短的持仓后卖出的比例占总卖出比例的 5%，六个交易点的级之间的相对变化具有非显著性，根据此结果拟合参数，此时求得 $a=0.35,\beta=3.48,\gamma=0.52$。

4.4 波动价格短期预测模型

文章选取短期（以分钟为单位）和长期预测（以天为单位），对橡胶期货价格的波动建立预测模型。

4.4.1 自回归与分布滞后模型

本文借鉴论文[16][17]中的对价格滞后模型的研究，建立价格波动与成交量，持仓量关系。得到如下模型

$$\hat{\sigma}_{\mathrm{t}}=\mu+\sum_{i=1}^{m}\rho_i\hat{\sigma}_{t-i}+\alpha NL_{\mathrm{t}}+\beta NC_t+\varepsilon_t$$

其中，$\hat{\sigma}_{\mathrm{t}}$ 为单位时间价格波动估计；NG_t 和 NC_t 分别为交易量和持仓量。

价格波动为 $German-Klass$ 波动估计量，计算如下：

$$\hat{\sigma}_t=\{0.5\times(\ln((P_{t,h}/P_{t,l}))^2-(2\ln(2)-1)(\ln(P_{t,o}/P_{t,c}))^2\}^{1/2}$$

其中 $P_{t,h}$ 和 $P_{t,l}$ 为单位时间成交价极大和极小值，$P_{t,o}$和$P_{t,c}$ 为单位时间的开盘价和收盘价。

本文将单位时间价格波动看做单位时间成交价的价差。

$$\hat{\sigma}_t=NJ_t-NJ_{t-1}$$

同时考虑因变量对结果的滞后效应引入多元分布滞后模型：

$$\hat{\sigma}_t=\mu+\sum_{i=1}^{m}\rho_i\sigma_{t\text{-}i}+\sum_{j=1}^{n}\alpha_j NL_{\mathrm{t\text{-}j}}+\sum_{k=1}^{l}\beta_k NC_{t-k}+\varepsilon_t$$

将公式累加：

$$NJ_t=\mu+\sum_{i=1}^{n}\rho_i NJ_{t-i}+\sum_{i=1}^{n}\rho_i NZ_{t-i}+\sum_{i=1}^{n}\rho_i NY_{t-i}+\varepsilon_t$$

选取 1 分钟为单位时间，对橡胶期货交易进行统计分析，对当月其他时间段的期货成交价（NJ）进行预测。数据处理：成交价采用每个单位时间的平均成交价，成交量采用单位时间的成交总量（NZ），持仓量采用单位时间的持仓总量

(NY)。

- 用 Matlab 做出自变量与因变量的散点图

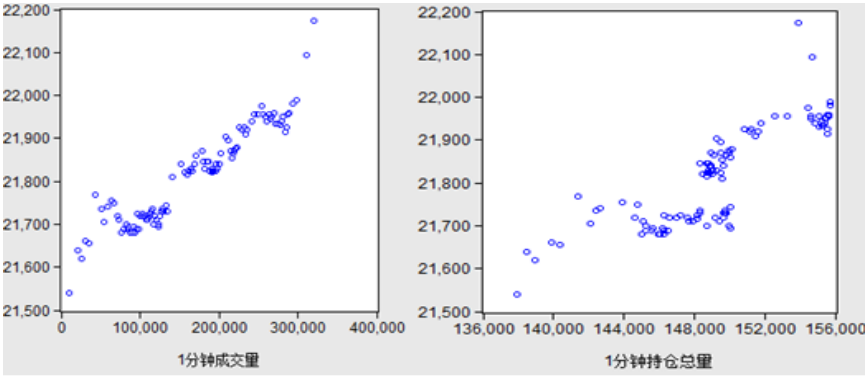


图 15 自变量与因变量散点图

从散点图可以看出具有一定的线性关系，即可选用每分钟的成交量（NL）与持仓（NC）量作为自变量。

- 用 Eviews 软件对因变量成交价（NJ）进行高阶自相关性检验，编程计算得到偏相关系数，变量间的相关程度。数据如下

表 13 高阶自相关检验

滞后期	自相关系数	偏相关系数	Q-Stat	Prob
1	0.894	0.894	87.229	0.000
2	0.830	0.150	163.08	0.000
3	0.787	0.111	231.94	0.000
4	0.755	0.082	231.94	0.000

偏相关系数>0.5,可以得到成交价 NJ 有 1 阶自相关性。

- 估计自相关滞后阶数及自变量滞后阶数（在第一问中通过格兰杰因果检验对自变量的滞后性进行了说明）

用 Eviews 软件估计自相关滞后系数为 1，自变量 NJ 滞后系数为 16，NC 滞后系数估计为 10。在软件中编代码进行模型估计，要兼顾 P 值的大小，最后取滞后系数分别为 8,3。得到模型：

$$NJ_t = \mu + \sum_{i=1}^{m} \rho_i NJ_{t-i} + \sum_{j=1}^{n} \alpha_j NZ_{\text{t-j}} + \sum_{k=1}^{l} \beta_k NY_{t-k}$$

经计算得到

$$\alpha_{\text{j}} = [0.00057, 0.00044, 0.00031, 0.00018, 4.5E-5, -8.5E-05, -0.00021, -0.00034, -0.00047]$$

$$\beta_k = [-0.00585, -0.00209, 0.00167, 0.00543] \quad \rho_i = [0.763664] \quad \mu = 5201.369$$

检验结果如下

表 14 模型检验表

R-squared	F-statistic	Prob(F-statistic)	S.D. dependent var
0.961841	463.7990	0.000000	107.0749

数据显示拟合程度较好。

- 输入数据得到变量 NJ 的拟合曲线及数据散点图的吻合情况并如图 16

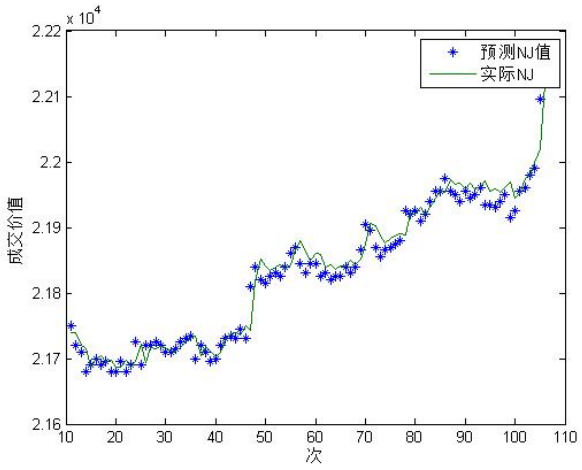


图 16 拟合曲线与散点图

对下午的价格波动进行预测，如图 17

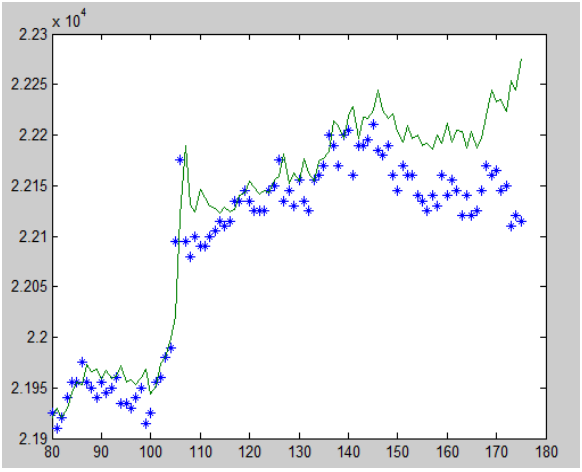


图 17 预测曲线与散点图

从预测可以看出，在下午的初期拟合程度较好，接下来一段时间偏离很大。因此说明该模型的建立对短期的预测是有帮助的，同时模型的参数需要随着时间不断的修正。

4.4.2 小波神经网络预测模型

小波分析是针对傅利叶变换的不足发展而来的，傅利叶变换有一个严重不足，就是变换抛弃了时间信息。小波是一种长度有限、平均值为 0 的波形

$$f_{\Gamma}(a,\tau)=\frac{1}{\sqrt{a}}\int_{t_1}^{t_2}x(t)\phi(\frac{t-\tau}{a})dt,a>0$$

τ相当于使镜头相对于目标平行移动的，a相当于使镜头向目标推进或远离。

小波神经网络[18][19][20]是一种 BP 神经网络拓扑结构为基础，把小波函数作为隐含层节点的传递函数，信号前向传播的同时误差反向循环播的神经网络。拓扑结构如图：

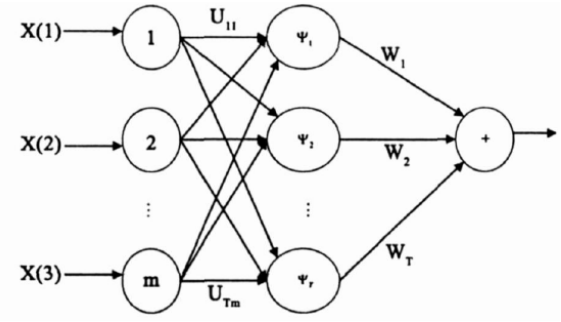


图 18 小波神经网络图

$X_1, X_2, ..., X_k$ 是输入参数，$Y_1, Y_2, ..., Y_m$ 是预测输出，U_{TM} 和 W_T 为权值。在输入信号序列为 $x_i(i=1,2,...,k)$ 时，隐含层输出计算公式为

$$h(j) = h_j\left(\frac{\sum_{i=1}^{k} \omega_{ij} x_i - b_j}{a_j}\right), j = 1,2,...,l$$

本题采用小波基函数为 Morlet 母小波基函数，数学公式为

$$y = \cos(1.75x)e^{\frac{x^2}{2}}$$

小波神经网络输出层计算公式为

$$y(k) = \sum_{i=1}^{i} \omega_{ik} h(i), k = 1,2,...m$$

算法训练步骤如下：

step1：网络初始化。

step 2：样本分类。

step 3：预测输出。

step 4：权值修正。

step 5：判断算法是否结束，若没有结束，返回 step 3。

根据小波分析的方法，以影响价格的 3 个因素作为输入，对它们分别进行初始化，训练、修正。并对预测结果进行分析。得到迭代 200 次预测结果图 19。

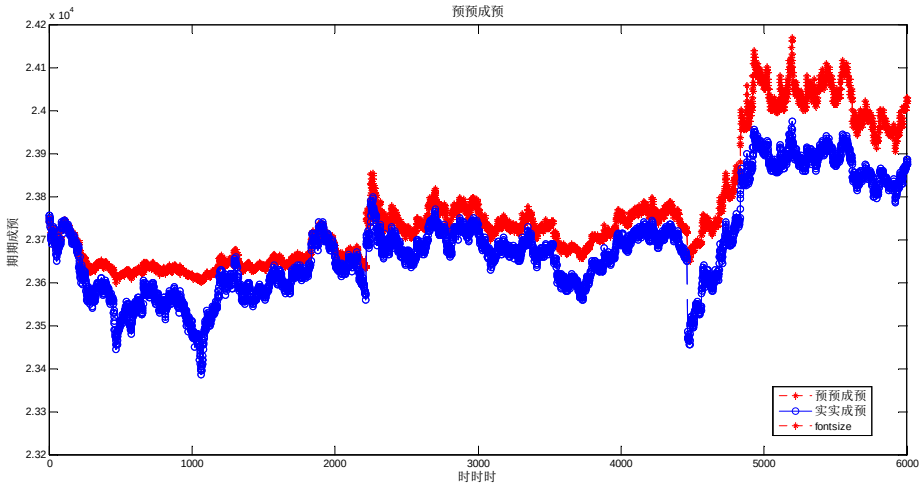


图 19 模型训练结果

由上图可知，训练 200 次后的小波预测结果已大致描述出了价格的趋势，

为了进一步精确预测的结果，迭代 500 次的训练，预测结果如图 20。

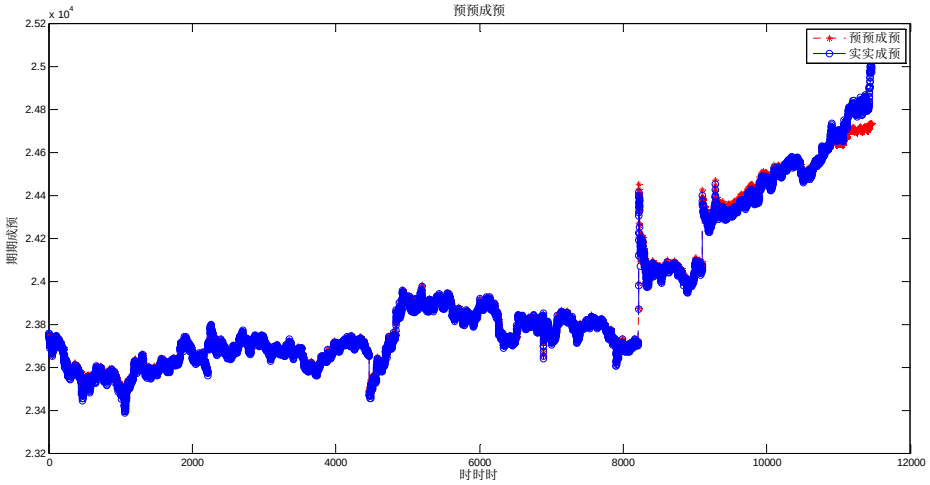


图 20 小波神经网络训练结果

由上图可见，训练 500 次以后的小波进行预测的结与实际情况的吻合度较高。预测结果的平均相对误差 0.627，预测结果的精确度达到 0.9493。小波神经网络的预测结果已经相当精确。故该小波预测价格模型符合要求。

4.5 波动价格长期预测

以一天为单位时间，对 3 至 28 日的模型进行分析，得到单位时间的 *NC* 、 *NL* 、 *NS* 、 *NBL* 处理后的数据，其中成交价采用单位时间成交价的均数，成交量采用单位时间末的成交总量，持仓量采用单位时间末的持仓量。

a.自变量与因变量之间以及其一阶差分之间的散点图：

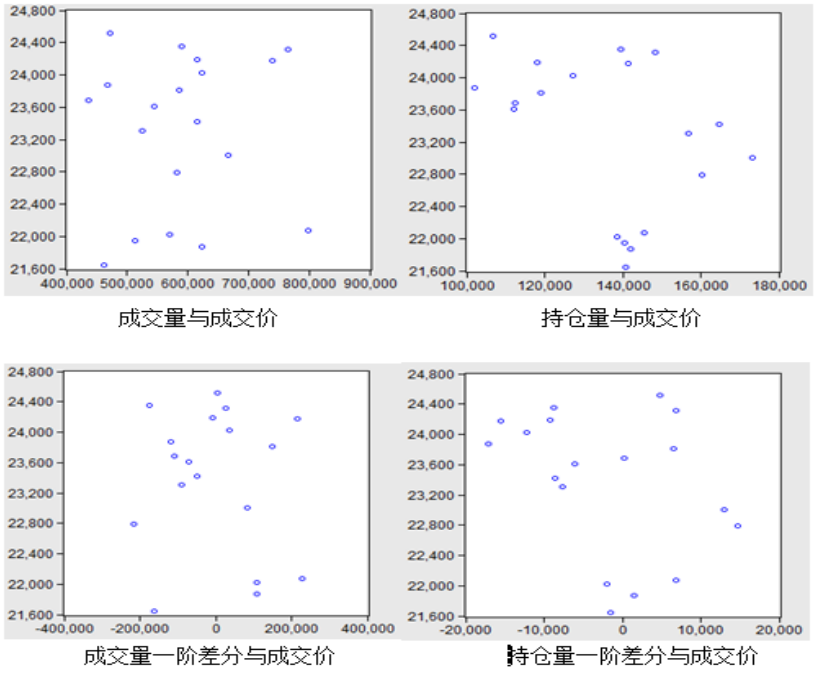


图 21 自变量与因变量之间的散点图

从上图看不出其相关关系，不能用短期预测的方法，所以可以采用时间序列方法

对价格变动进行分析和预测。

b.用 sas 软件对成交价进行时间序列分析

考虑数据的趋势性和滞后性，选择 10 个模型进行模拟，自动选择出最符合要求的模型。最后得到阻尼衰减趋势指数平滑模型。指数平滑即对历史的加权平均，可以用在没有任何一种明细函数规律，有着某种前后关联的时间序列的短期预测，其特点是可以跟踪数据变化，在预测中，会添加新的样本数据，新数据取代老数据的位置，因而预测值能够反映最新的数据结构。

采用 3 日至 26 日的数据模拟，对 27 日和 28 日的价格进行预测，将预测值与真实值数据进行对比：

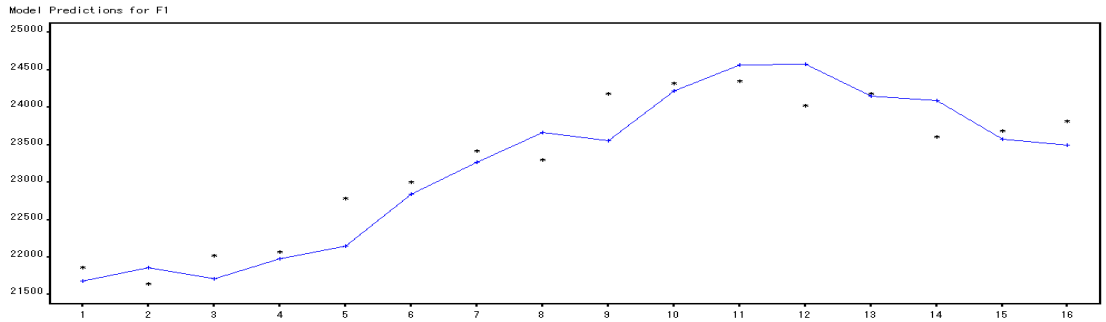


图 22　价格波动拟合图

表 15 拟合统计量

均方差	标准误差	绝对平均误差	相关系数
116606.3	341.47667	251.07446	0.865

表 16 预测结果分析

时间	2012 年 9 月 27 日	2012 年 9 月 28 日
预测值	23909	24075
实际值	23879.11	24518.79
中位数	23815	24500
偏度	0.430249	0.462597674
极差	820	820

实际价格中位数与平均数相差不大，成交价偏度的绝对值较小，说明数据在中位数的两侧大小分布较均匀，因而成交价的平均数能反映一天的实际价格，在预测之中 9 月 27 号的成交价与预测值相差为-29.89，极差的比值为 3%，预测情况较好，而在 9 月 28 号的预测值中，与成交价偏差较大，极差的比值为 54.12%，偏差很大，因而模型对波动较大的数据预测精度不足。

对比分析 9 月 28 号的成交价，成交量，持仓量变化。不难发现价格有上涨的趋势，而成交量在后期变大，持仓量减小，说明卖空和买空者都在大量平仓，价格马上会下跌。

4.6 投资交易模型

4.6.1 投资交易模型的建立

●目标函数

以模型二预测的价格 P_{yi} 与此时的成交价 P_{ci} 的对数差作为收益率 R_i（定义的依据是取对数之后做差与直接做差的收益率差别不大）

$$R_i = \ln P_{yi} - \ln P_{ci}$$

收益总和$_{TR}$ 与第i天买入量B_i，第 i 天卖出量C_i以及H_i、L_i之间的关系（H_i、L_i都是随着收益率变化的量，只取 0 和 1 数值）

$$TR = \sum_{i=1}^{\infty} f(B_i, C_i, H_i, L_i)\ i = 1,2,...,\infty$$

要确保投资者收益最大，对模型进行修正，结果如下

$$TR_{\max} = \sum_{i=1}^{\infty}[(P_{ci} * L_i * C_i - 20 * C_i) - (P_{ci} * H_i * B_i + 20 * B_i)] - 100\text{万}\ i = 1,2,...,\infty$$

令$Z_i = (P_{ci} * L_i * C_i - 20 * C_i) - (P_{ci} * H_i * B_i + 20 * B_i)\ i = 1,2,...,\infty$

则有$TR_{\max} = \sum_{i=1}^{\infty} Z_i\ i = 1,2,...,\infty$

●约束条件分析

H_i、L_i都是随着收益率变化的，只取 0 和 1 数值，约束条件：

$$H_i = \begin{cases} 1 & R_i > 0 \\ 0 & R_i < 0 \end{cases}\ i = 1,2,...,\infty$$

$$L_i = \begin{cases} 1 & R_i < 0 \\ 0 & R_i > 0 \end{cases}\ i = 1,2,...,\infty$$

期货交易存在杠杆效应，即在保证金制度下放大交易额，因此，风险也被放大。第四组约束条件依据杠杆效应：要求投资者每天的交易数额有不应该超出前天收盘时投资者所存的投资金额，如下：

$$F_i + (-1)^{H_i} * Z_i = F_{i+1} \qquad i = 1,2,...,\infty$$

$$10\% * Z_i < F_{i-1} \qquad i = 2,3,..\infty$$

$$F_1 = 100\text{万}$$

简化模型$TR_{\max} = \sum_{t=1}^{n}\sum_{i=1}^{\infty} Z_i$

约束条件：

$$H_i = \begin{cases} 1 & R_i > 0 \\ 0 & R_i < 0 \end{cases}\ i = 1,2,...,\infty$$

$$L_i = \begin{cases} 1 & R_i < 0 \\ 0 & R_i > 0 \end{cases}\ i = 1,2,...,\infty$$

$$F_i + (-1)^{H_i} * Z_i = F_{i+1} \qquad i = 1,2,...,\infty$$

$$10\% * Z_i < F_{i-1} \qquad i = 2,3,..\infty$$

还要结合价格波动模型，预测出价格变化趋势，通过H_i、L_i的 0 ，1 取值，来确定是开多单，还是做空，以及确定最佳时期平仓，从而减小风险，获取更大收益。

4.6.23 进制编码遗传算法的求解

将每个交易点看做一个个体的编码位，利用遗传算法求解，仅对预测的 30 个交易点选择优化，交易买卖为一一映射，同理，也可将更多的预测交易点纳入交易模型，基本算法步骤如下：

Step1：随机初始化种群，交易点即个体的选择一定，每个个体表示为 3 进制基因编码。0 表示无交易，1 表示买入，2 表示卖出。初始化和选择的个体必须满足条件，在每个交易点总买一手次数≥总卖一手次数,否则重新生成。

Step2：用轮盘赌策略确定个体的适应度即买卖一手的差价，并判断是否符合优化准则，若符合则输出最佳个体代表的最优解，并计算结果，否则转向 Step 3。

Step3：根据适应度选择再生个体，适应度高的个体被选择的概率高，适应度低的个体被选择的概率低，采用顺序选择遗传算法。

$$p_j = \frac{q(1-q)^{j-1}}{1-(1-q)^{NP}}$$

Step4：按照一定的交叉、变异和交叉、变异方法生成新个体

Step5：由交叉和变异生成新代种群，返回 Step2。

- 根据上文的结果，在一个周期内的买卖一手的持仓时间概率分布，选择每个周期的买卖次数时间点分别为 5 次。其余的时间点均未发生有效交易。利用 Matlab 的仿真的一次最优（编码为 10000210110000000200002200012）收益为 1340。交易的最优买入点策略：1 7 9 10 29 交易的最优卖出点策略：6 19 24 25 30。
- 考虑更为一般化的交易，买卖期货时不仅考虑一一映射，可以为一次多买入，多买入的一次卖出。同时根据期货持仓时间长度的概率分布约束每次的买卖出时长。
- 与实际交易的价格对比，实际价格的最优交易收益为 1463。收益的精度为 92.2%，收益的精度应该解释为小波神经网络的预测精度的问题。

分别选取数量为 30、40、50 的交易点长度和不同数量的买卖次数的收益精度。如表 17：

表 17：交易策略收益精度

交易点数量买卖次数	4	5	6
30	92.6%	93.7%	94.7%
40	91.9%	92.2%	93.6%
50	90.2%	93.1%	98.1%

五、结论与建议

本文在对实时期货交易数据处理时，首先通过散点图和相关系数矩阵，对影响期货价格的因素进行选取。通过合理的统计方法检验后可知：(1) 变量之间直接的相关性比较强，存在着信息上的重叠。(2) 橡胶期货价格的主要影响因素有 NJ,NL 和 NC。(3) 影响因素的一阶差分是平稳的时间序列。所以本文在进行价格研究时避免了直接根据经验选择的主观影响。

本文的交易价周期与分类具有以下优点：(1) 多周期并存，通过海量交易价的峰谷值分析法和傅里叶级数周期法，分别得出短中长三种周期，使得交易投资具有不同的选择性。(2) 期货的波动分类通过 R 型聚类分析，筛除冗余分类。(3) 对每种分类给出了期货的市场表现。

考虑到滞后影响，采用自回归与分布滞后模型与小波神经网络算法结合，对进行价格短期预测。(1) 对历史数据进行充分挖掘，对一些市场现象和变化进行解释。(2) 具有较高的短期预测精度。(3) 长期预测的具有较大的困难性。

在最优化交易模型中，利用遗传算法得到 NP 问题近似最优的多次买卖一手的交易策略，未能详尽的分析不同的交易策略。

本文也存在一些需要后续继续研究的问题。重点是通过对橡胶交易大数据进行周期与波动分析，对我国期货的价格预测提供一种新的思路，由于时间、样本信息及个人能力的限制，未能对其他因素的影响作分析。例如，价格模型中未考虑到的季节变化，这一定程度上影响了预测模型的长期预测的可能性。

六、参考文献

[1]翟光磊. 橡胶期货价格和交易量、持仓量的相关性分析[J]. 金融经济,2011,18:75-77.

[2]包娟.中国大豆期货市场和现货市场的动态关系研究[D].江西财经大学,2009.DOI:10.7666/d.y1658288.

[3]陈昆亭,周炎,龚六堂. 中国经济周期波动特征分析:滤波方法的应用[J]. 世界经济,2004,10:47-56+80.

[4]桂俊煜. 我国天然橡胶期货市场与现货市场价格关系的研究[D].北京林业大学,2013.

[5]卢垚. 中国天然橡胶期货与现货价格关系实证研究[D].北京大学,2007.

[6]中国期货业协会官方网站(www.cfachina.org).

[7]田新民,沈小刚. 基于交易量和持仓量的期货日内价格波动研究[J]. 经济与管理究,2005,07:78-80.

[8]王常亮. 基于三种滤波方法的中国经济周期波动特征研究[D].山东大学,2012.

[9]戴胜利,童身以. 周期图法在疾病预测中的应用[J]. 中国卫生统计,1988,03:45-46.

[10]陈昆亭,周炎,龚六堂. 中国经济周期波动特征分析:滤波方法的应用[J]. 世界经济,2004,10:47-56+80.

[11]苗敬毅. 中国股市波动性的隐周期研究[J]. 数理统计与管理,2008,05:905-910.

[12]郝杰. 基于改进小波神经网络的上证指数预测研究[D].华南理工大学,2014.

七、附录

附件一：原数据、预处理后数据、进行数据分析预测的数据

附件二：小波神经网络预测、滤波分析、分布滞后预测、3进制遗传算法。