재현그림을 통한 우리나라 환율 자료에 대한 탐색적 자료분석

장대홍¹

¹부경대학교 통계학과

접수 2013년 6월 30일, 수정 2013년 7월 25일, 게재확정 2013년 8월 2일

요 약

탐색적 자료분석에서는 자료를 통계적 모형에 바로 적합시키기 보다는 자료를 있는 그대로 보려는 데 주안점을 둔다. 우리는 시계열 자료에 대한 그래픽 탐색적 자료분석방법의 하나로써 재현그림을 사용할 수 있다. 재현그림의 장점은 통계모형에 대한 가정 없이 시계열 자료의 구조적 패턴을 확인할 수 있고 이 패턴을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있다는 데 있다.

주요용어: 일반화자기회귀조건부이분산성모형, 자기회귀적분이동평균모형, 재현그림, 정상성, 환율데이터.

1. 서론

모형에 기반을 둔 시계열 자료에 대한 통계적 분석으로서 자기회귀적분이동평균모형 (ARIMA 모형), 자기회귀조건부이분산성모형 (ARCH 모형), 일반화자기회귀조건부이분산성모형 (GARCH 모형) 등을 주로 사용한다 (Lee와 Kim, 2006; Shim과 Lee, 2010). 탐색적 자료분석에서는 자료를 모형에 바로 적합시키기 보다는 자료를 있는 그대로 보려는 데 주안점을 둔다. 시계열 자료에 대한 탐색적 자료분석방법으로서 우리는 주로 단순평활기 (이동중앙값 (3, 4, 5), 해닝 (H), 분리 (S), 반복 (R) 등)와 복합평활기 (4253H, 3RSSH, 3RS3R 등)를 사용한다. 이 평활기들은 잡음이 섞인 시계열 자료에서 신호를 매끄럽게 분리하기 위한 탐색적 방법이다.

Eckmann 등 (1987)은 동적 시스템의 위상공간 (phase space) 상태의 재현을 시각화하는 재현그림 (recurrence plot)을 제안하였다. 재현그림은 이론물리학에서 비선형 시계열이나 혼돈 시스템의 패턴과 성질을 밝히기 위하여 고안되었다. Marwan 등 (2007)은 재현그림에 대한 리뷰 페이퍼를 잘 정리하여 발표하였다.

이 재현그림은 최근에 다양한 학문분야 (경제/경영학, 음향학, 생물학, 지질학, 의학, 전산학 등)에서 응용되고 있다. 국내 응용연구로는 Jang (2009)이 난수의 임의성을 평가하기 탐색적 그림도구로서 재현그림을 적용하였고 Jang (2013)은 재현그림을 통한 우리나라 주식자료에 대한 탐색적자료분석을 행하였다. 국외 연구로는 경제/경영학 분야에서 Belaire-Franch 등 (2002) 이 재현그림을 통한 환율 자료의 분석을 시도하였고 재현그림을 통한 주식시장에 대한 분석으로서는 Guhathakurta 등 (2010)이 내생 주식시장 붕괴와 외생 주식시장 붕괴를 구별하기 위한 수단으로서 재현그림과 재현수량화분석 (RQA; recurrence quantification analysis) 을 사용할 것을 제안하였다. Chen (2011)은 재현그림과 재현수량화분석을 사용하여 대만의 실업률 데이터에 대한 분석을 행하였다. Addo 등 (2013)은 재정위기를 탐지하기 위한 수단으로서 재현그림을 제시하였다.

¹ (608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과, 교수. E-mail: dhjang@pknu.ac.kr

우리는 시계열 자료값 자체 보다 시계열 자료값의 차이에 관심을 가질 때가 종종 있다. 예로 주가분석이나 환율분석에서 주가나 환율의 차이가 중요한 경우가 있다. 이 때 재현그림이 시계열 자료에 대한 탐색적 그림도구로서 유용하게 쓰일 수 있다. 이러한 재현 그림을 구현하는 데 있어서 통계모형에 대한 가정이 필요 없으므로 그래픽 탐색적 자료분석 도구로서 유용한 수단이 될 수 있다.

Kim과 Kim (2013)은 우리나라 환율 원/달러 데이터를 사용하여 모형적합을 시도하였고 ARIMA(2, 1, 2) 모형과 ARIMA(1, 1, 0)+IGARCH(1, 1) 모형을 제시하고 이 두 모형들을 이용하여 구조 변화점을 추정하였다. 사용한 환율 데이터는 1995년 1월 3일부터 2012년 1월 31일까지 17년간 일별 환율 데이터로서 총 데이터 개수는 4,615개이며 외환은행 환율고시의 매매기준율 (basic rate of exchange)를 사용하였고 단위는 원/달러이다. 본 논문에서는 Kim과 Kim (2013)의 결과와 비교하기 위하여 같은 환율 데이터를 대상으로 시계열 자료값과 자료값의 차이에 초점을 맞추어 재현그림을 작성하고 시계열 데이터에 대한 패턴을 살폈다. 또한 본 논문의 연구결과를 Kim과 Kim (2013)의 결과와 비교하여 보았다.

2. 재현그림

시계열자료 $u_t (t = 1, 2, \dots, N_0)$ 가 주어졌을 때 위상공간은 테이켄스정리 (Takens' theorem)에 의하여 다음과 같이 재구성될 수 있다.

$$\mathbf{x}_i = (u_i, u_{i+\tau}, \dots, u_{i+\tau(m-1)})$$

여기서 $\mathbf{x}_i, i = 1, 2, \dots, N$ 은 위상공간 궤도 (trajectory)이고, N 은 위상공간 상태의 개수, u_i 는 시계열이고 m 은 매립차원 (embedding dimension)이며 τ 는 시간지연 (time delay)이고 $N_0 = N + \tau(m-1)$ 이다. 재현그림을 통하여 우리는 m -차원 위상공간 궤도를 2차원으로 표시하여 탐구할 수 있다. (분계)재현그림 ((thresholded) recurrence plot)은 다음과 같은 재현행렬로 표현할 수 있다.

$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (2.1)$$

여기서 $\mathbf{x}_i, \mathbf{x}_j \in R^m, i, j = 1, 2, \dots, N$ 이고, ϵ 은 분계점 (threshold) 거리, $\|\cdot\|$ 은 노름, Θ 는 헤비사이드 (Heaviside) 함수이다. 헤비사이드 (Heaviside) 함수 $\Theta(x)$ 는 $x < 0$ 일 때는 0이고 $x \geq 0$ 일 때는 1이 되는 함수이다. 두 개의 위상공간 $\mathbf{x}_i$ 와 $\mathbf{x}_j$ 의 거리가 ϵ 보다 작거나 같으면 위상공간 궤도가 같은 위치에 있다고 보아 $R_{i,j}(\epsilon) = 1$ 이 되어 정사각형 모양의 재현그림에서 x 축에서의 i -번째 시점과 y 축에서의 j -번째 시점에 해당하는 위치에 점으로 나타나게 된다. 모든 $i, j = 1, 2, \dots, N$ 에 대하여 $R_{i,j}(\epsilon) = 1$ 일 때만 (i, j) 위치에 점을 찍으면 재현그림이 완성된다.

재현그림에 나타나는 점들의 패턴을 살펴봄으로써 시계열 자료값의 재현성을 탐색적으로 검토하여 볼 수 있다. 환율분석의 예에서 재현그림 상 두 시점에서 점이 나타난다면 이 두 시점에서의 환율이 거의 유사하다는 것을 알 수 있게 되고 재현 그림상의 집락이나 패턴을 통하여 어떤 시점들의 그룹이 어떤 시점들의 그룹과 연관성을 갖는 지를 탐색적으로 알 수 있게 된다.

재현그림에서 중요한 모수가 분계점 ϵ 의 값이다. ϵ 이 너무 작으면 정사각형의 재현그림에서 점들을 거의 볼 수 없게 되어 빈 공간에 가까운 재현그림이 되고, ϵ 이 너무 크면 재현그림에서 거의 모든 점들이 다른 점들의 이웃이 되어 정사각형의 공간을 모든 점들이 채워 거의 까맣게 되는 재현그림이 된다. Mindlin과 Gilmore (1992), Zbilut와 Webber Jr. (1992), Matassini 등 (2002), Thiel 등 (2002), Zbilut 등 (2002)이 분계점 ϵ 의 값을 선택하는 방법들에 대하여 연구하였다.

재현그림과 더불어 우리는 비분계재현그림 (unthresholded recurrence plot)을 그릴 수 있는 데 비분계그림은 다음과 같은 거리행렬로 표현할 수 있다.

$$D_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\| \quad (2.2)$$

여기서 $\mathbf{x}_i, \mathbf{x}_j \in R^m, i, j = 1, 2, \dots, N$ 이고, $\|\cdot\|$ 은 노름이다.

어떻게 재현그림을 시계열 자료에 대한 탐색적 그림도구로서 쓸 수 있는 지를 우리나라 환율 데이터를 이용하여 살펴볼 수 있다. Iwanski와 Bradley (1998)는 저차원 동적 시스템에서 재현그림 통계량들 ($DET, LAM, L_{max}, V_{max}, L, TT$ 등)이 매립차원 m 에 불변 (invariant)함을 보였다. 이런 이유와 더불어 탐색적 자료분석의 성격상 매립차원 m 과 시간지연 τ 를 고려할 필요가 없이 원 시계열 자료 자체를 이용하면 된다. 즉 $m = 1$ 로 놓으면 된다. 또한 노름은 유클리드 거리를 이용하기로 하고 모든 예제에서 특별한 언급이 없는 한 ϵ 으로 시계열 자료에 대한 범위의 1/10배 또는 1/20배를 이용하기로 한다. 재현그림을 그리기 위한 소프트웨어로서 Visual Recurrence Analysis와 CRP를 사용하였다.

Figure 2.1은 환율 데이터와 환율 데이터에 대한 차분 데이터를 나타낸다. 1차 차분을 통하여 평균이 일정하여 졌음을 알 수 있다. Kim과 Kim (2013)은 ADF 검정을 통하여 환율 데이터가 비정상 시계열임을 보이고 환율데이터에 대한 차분 데이터가 정상 시계열이 됨을 보였다. Figure 2.2는 차분 데이터에 대한 바이올린그림, 상자그림 그리고 정규확률플롯이다. 특이값들이 다수 존재함을 알 수 있다. Shapiro-Wilk 검정을 행하니 정규성으로부터 많이 벗어남을 알 수 있다 (p -값 < 0.0001). 즉 정규분포보다 꼬리가 매우 두꺼움 (왜도가 -0.2806, 첨도가 208.2110임)을 알 수 있어 변동성이 큼을 알 수 있다.

Figure 2.1 Exchange rate data and first differenced exchange rate data

Figure 2.2 Violin plot, box plot, and normal probability plot for first differenced exchange rate data

3. 재현그림을 통한 탐색적 자료분석

환율 데이터에 대하여 비분계재현그림을 그려보면 Figure 3.1과 같다. 칼라색이 골고루 퍼져 있지 않고 다수의 특정 색값들의 클러스터를 확인할 수 있어 비정상 시계열임을 확인할 수 있다. Kim과

Kim (2013)은 ARIMA(2,1,2) 모형과 ARIMA(1,1,0) + IGARCH(1,1) 모형들을 이용하여 구조 변화점을 추정하여 ARIMA(2,1,2) 모형에서는 $t=862, 1748, 2815, 3766$ 를 제시하였고 ARIMA(1,1,0) + IGARCH(1,1) 모형에서는 $t=882$ 를 제시하였다. Figure 3.1의 비분계재현그림의 공간패턴은 크게 3등분 (3등분되는 구간을 구간 1, 2, 3라고 하자.)되며 이 3등분되는 위치는 $t=868, 3764$ 이어서 Kim 과 Kim (2013)이 찾은 중요한 두 개의 구조 변화점 위치 ($t=862, 3766$)와 아주 비슷함을 알 수 있다. 즉 1997년 IMF 외환위기와 2008년 미국 서브프라임사태로 촉발된 금융위기에 해당한다. Figure 3.1의 비분계재현그림 공간패턴을 보면 구간 2는 다시 3 등분 (구간 2-1, 2-2, 2-3이라 하자.)됨을 알 수 있고 이 소구간들이 3등분되는 위치가 Kim과 Kim (2013)이 찾은 또 다른 두 개의 구조 변화점 위치 ($t=1748, 2815$)와 아주 비슷함을 알 수 있다. 그러므로 우리는 재현그림을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있게 된다.

Figure 3.1의 비분계재현그림 공간패턴을 다시 살펴보면 구간 2가 2등분된다고 볼 수도 있다. 즉 Kim과 Kim (2013)이 찾은 구조 변화점 위치 ($t=2815$)를 중심으로 구간 2가 2등분된다.

Figure 3.1 Unthresholded recurrence plot for exchange rate data

Figure 3.2는 환율 데이터에 대한 (분계)재현그림으로 점들이 골고루 퍼져 있지 않고, 특히 좌상, 우하 구석의 빈 공간 및 중간중간의 빈 공간으로 인하여 비정상 시계열임을 확인할 수 있다. 또한 Chen (2011)이 제시했던 P-C-L-C 싸이클과 유사한 싸이클을 확인할 수 있고 이 싸이클을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있게 된다. P-C-L-C 싸이클이란 시계열 처음에서 P (periodic zone)이 나타나고 다음에 차례로 C (chaotic-looking zone), L (laminar patch), C (chaotic-looking zone)가 뒤따른다는 것을 말한다. P는 환율의 주기성이 나타나는 영역, C는 환율의 급락 또는 급등이 나타나는 영역, 그리고 L은 환율의 변동이 심하지 않은 안정된 영역이다. Figure 3.2의 재현그림에서 PL-C 싸이클이 반복해서 나타난다. 여기서 PL이란 P (periodic zone) 와 L (laminar patch)이 혼재되어 있는 영역을 가리킨다. 환율 데이터에서는 크게 보면 PL-C 두 싸이클 다음에 PL로 마감된다. 두 개의 C는 1997년 IMF 외환위기와 2008년 미국 서브프라임사태로 촉발된 금융위기에 해당한다.

Figure 3.2 (Thresholded) recurrence plot for exchange rate data

우리는 재현수량화분석을 통하여서도 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있게 된다. 재현그림에서 점들의 미세한 구조는 다음 Table 3.1과 같은 재현수량화분석 측도값들을 사용하여 복잡성 (complexity)을 측정할 수 있다.

Table 3.1 Measures in recurrence quantification analysis

Measures	Names	Contents	Numeric Equations
Variance	Variance	variance	
RR	Recurrence Rate	the density of recurrence points	$\frac{1}{N} \sum_{i,j=1}^N R_{i,j}(\epsilon)$
DET	Determinism	the ratio of recurrence points that form diagonal structures (of at least length l_{min}) to all recurrence points	$\frac{\sum_{l=l_{min}}^N lP(l)}{\sum_{l=1}^N lP(l)}$, where $P(l)$ is the density function of diagonal lines of length l
L	Length	the average diagonal line length	$\frac{\sum_{l=l_{min}}^N lP(l)}{\sum_{l=l_{min}}^N P(l)}$
ENTR	Entropy	Shannon Entropy of $p(l)$ to find a diagonal line of exactly length l	$-\sum_{l=l_{min}}^N p(l) \ln p(l)$, where $p(l) = P(l)/N_l$, $N_l = \sum_{l \geq l_{min}} P(l)$
LAM	Laminarity	the ratio of recurrence points that form vertical structures (of at least length v_{min}) to all recurrence points	$\frac{\sum_{v=v_{min}}^N vP(v)}{\sum_{v=1}^N vP(v)}$, where $P(v)$ is the density function of vertical lines of length l
TT	Trapping Time	the average vertical line length	$\frac{\sum_{v=v_{min}}^N vP(v)}{\sum_{v=v_{min}}^N P(v)}$
RTE	Recurrence Time Entropy	the measure for determining the periodicity	$-(\ln T_{max})^{-1} \sum_{t=1}^{T_{max}} P(t) \ln P(t)$, where $P(t)$ is the recurrence period density function and T_{max} is the largest recurrence value
T_2	Recurrence Times of Second Type	recurrence times of second type	$T_j^2 = \{i, j : \mathbf{x}_i, \mathbf{x}_j \in R_i; \mathbf{x}_{j-1} \notin R_i\} $, where R_i are the recurrence points which belong to the state $\mathbf{x}_i$

Figure 3.3은 창 (window)의 크기를 100으로 하여 창을 1에서 시작하여 1씩 증가하여 움직이며 구한 각 종 재현수량화분석을 위한 측도값들을 그린 그림들이다 (대각선 최소길이 l_{min} 과 수직선 최소길이 v_{min} 는 2로 하였음). PL-C 사이클을 확인할 수 있고 이 사이클을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있다. 환율 데이터에서 PL-C 두 사이클 다음에 PL로 마감됨을 볼 수 있다. 분산값을 보면 두 개의 C, 즉 1997년 IMF 외환위기와 2008년 미국 서브프라임사태로 촉발된 금융위기에서 변동성이 크게 증가함을 알 수 있다. 특히 첫 번째 C (1997년 IMF 외환위기)에서 변동성이 매우 커졌음을 알 수 있다. $RR, DET, L, ENTR, LAM, TT$ 에서는 깊은 골짜기를 통하여, RTE 에서는 피크를 통하여, T_2 에서는 값의 심한 변동을 통하여 두 개의 C를 확인할 수 있다.

Figure 3.3 RQA measures for exchange rate data (window size: 100)

구간 1, 2-1, 2-2, 2-3, 3은 각각 그 내부에 또 다른 작은 PL-C 사이클을 포함함을 알 수 있어 유사 자기상사성 (pseudo self-similarity) 현상이 나타남을 알 수 있다. 이러한 구조를 확인하기 위하여 전체 환율 데이터를 5개의 구간 (1, 2-1, 2-2, 2-3, 3)으로 나누어 각각의 재현그림 및 재현수량화분석 측도값들을 그려보면 5개의 각 구간에서도 PL-C 사이클이 존재함을 알 수 있다. 물론 PL-C 사이클의 형태는 구간에 따라 다르게 나타난다. 예를 들어 구간 2-2 데이터에 대한 재현그림과 각 종 재현수량화분석 측도값들을 그리면 Figure 3.4와 같다.

Figure 3.4 (a) Recurrence plot and (b) RQA measures (window size: 100) for interval (2-2) in exchange rate data

차분 데이터에 대하여 비분계재현그림과 (분계)재현그림을 그리면 Figure 3.5과 같고 재현수량화분석 척도값들을 그려보면 Figure 3.6과 같다. Figure 3.5을 통하여 점들이 상당히 골고루 퍼져 있어 차분 데이터는 정상 시계열이 됨을 확인할 수 있다. 비정상 시계열인 환율 데이터에 대한 재현그림과 비교하여 보면 상당히 다른 패턴을 지니고 있다. 그림에도 불구하고 재현그림을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있게 된다. PL-C 싸이클을 확인할 수 있고 이 싸이클을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있다. 환율 데이터에서처럼 PL-C 두 싸이클 다음에 PL로 마감됨을 볼 수 있다. 분산값을 보면 두 개의 C, 즉 1997년 IMF 외환위기와 2008년 미국 서브프라임사태로 촉발된 금융위기에서 변동성이 크게 증가함을 알 수 있다. 특히 첫 번째 C (1997년 IMF 외환위기)에서 변동성이 매우 컸음을 알 수 있다.

Figure 3.5 Unthresholded recurrence plot and recurrence plot for first differenced exchange rate data

Figure 3.6 RQA Measures for first differenced exchange rate data (window size: 100)

4. 결론

다양한 시계열 자료에 대한 그래픽 탐색적 자료분석 도구로서 재현그림을 사용하면 우리는 통계모형에 대한 가정 없이 재현그림을 통하여 시계열 자료의 구조적 패턴을 확인할 수 있고 이 패턴을 통하여 탐색적으로 시계열 데이터의 구조 변화점을 한 눈에 확인할 수 있게 된다. 또한 재현그림은 수식이 아닌 그림이므로 시계열 자료의 구조적 패턴을 시각적으로 확인할 수 있는 장점이 있다. 우리는 시계열 자료에 대한 확증적 자료분석 전에 보조적인 그래픽 수단으로서 재현그림을 활용할 수 있다.

References

- Addo, P. A., Billio, M. and Guegan, D. (2013). Nonlinear dynamics and recurrence plots for detecting financial crisis. *North American Journal of Economics and Finance*, in press.
- Belaire-Franch, J., Contreras, D. and Tordera-Lledo, L. (2002). Assessing nonlinear structures in real exchange rates using recurrence plot strategies. *Physica D*, **171**, 249-264.
- Chen, W. (2011). Use of recurrence plot and recurrence quantification analysis in Taiwan unemployment rate time series. *Physica A*, **390**, 1332-1342.
- Eckmann, J. P., Kamphorst, S. O. and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters*, **5**, 973-977.

- Guhathakurta, K., Bhattacharya, B. and Chowdhury, A. R. (2010). Using recurrence plot analysis to distinguish between endogenous and exogenous stock market crashes. *Physica A*, **389**, 1874-1882.
- Iwanski, J. S. and Bradley, E. (1998). Recurrence plots of experimental data: To embed or not to embed. *Chaos*, **8**, 861-871.
- Jang, D. H. (2009). Recurrence plots as an exploratory graphical tool for evaluating randomness. *The Korean Journal of Applied Statistics*, **22**, 1153-1165.
- Jang, D. H. (2013). Exploratory data analysis for Korean stock data with recurrence plots. *The Korean Journal of Applied Statistics*, **26**, 807-819.
- Kim, B. M. and Kim, J. H. (2013). Time series models for daily exchange rate data. *The Korean Journal of Applied Statistics*, **26**, 1-14.
- Lee, O. and Kim, M. J. (2006). Long memory and covariance stationarity of asymmetric power FIGARCH model. *Journal of the Korean Data & Information Science Society*, **17**, 983-990.
- Marwan, N., Romano, M. C., Thiel, M. and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, **438**, 237-329.
- Matassini, L., Kantz, H., Holyst, J. and Hegger, R. (2002). Optimizing of recurrence plots for noise reduction. *Physical Review E*, **65**, 021102.
- Mindlin, G. M. and Gilmore, R. (1992). Topological analysis and synthesis of chaotic time series. *Physica D*, **58**, 229-242.
- Shim, J. Y. and Lee, J. T. (2010). Estimation of nonlinear GARCH-M model. *Journal of the Korean Data & Information Science Society*, **21**, 831-839.
- Thiel, M., Romano, M. C., Kurths, J., Meucci, R., Allaria, E. and Arecchi, F. T. (2002). Influence of observational noise on the recurrence quantification analysis. *Physica D*, **171**, 138-152.
- Zbilut, J. P., and Webber Jr., C. L. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, **171**, 199-203.
- Zbilut, J. P., Zaldívar-Commenges, J. M. and Strozzi, F. (2002). Recurrence quantification based Liapunov exponents for monitoring divergence in experimental data. *Physics Letters A*, **297**, 173-181.

Exploratory data analysis for Korean daily exchange rate data with recurrence plots

Dae-Heung Jang¹

¹Department of Statistics, Pukyong National University

Received 30 June 2013, revised 25 July 2013, accepted 2 August 2013

Abstract

Exploratory data analysis focuses mostly on data exploration instead of model fitting. We can use the recurrence plot as a graphical exploratory data analysis tool. With the recurrence plot, we can obtain the structural pattern of the time series and recognize the structural change points in time series at a glance.

Keywords: ARIMA model, exchange data, GARCH model, recurrence plot, stationarity.

¹ Professor, Department of Statistics, Pukyong National University, Busan 712-749, Korea.
E-mail: dhjang@pknu.ac.kr