

# 밀도구분도 매핑을 이용한 통계정보 공간 내삽의 유효성 평가

## An Evaluation of Spatial Interpolation of Statistical Information Using Dasymetric Mapping

이병길<sup>1)</sup>

Lee, Byoung-Kil

### Abstract

For integrating and utilizing the statistical data, which is summarized by arbitrary areal unit such as demographics, with stellite imagery or other GIS data, areal unit of both data should be accorded. Dasymetric mapping is proposed as a useful method for disaggregating the aggregated statistical data to finer areal unit or generating surface model from object data such as polygonal area. This research evaluate the effectiveness of dasymetric mapping by 1) summarizing the yellow page information by administrative district, 2) modeling the business density using dasymetric mapping, and 3) comparing the business densities of raw data and that of spatial interpolation result.

Keywords : Dasymetric mapping, Areal Unit, Spatial interpolation

### 초 록

인구통계와 같이 임의의 공간 단위로 집계되는 통계자료를 위성영상이나, 여타 다른 GIS 도형정보와 통합하여 활용하기 위해서는 상호간의 공간 단위를 일치시켜야 한다. 밀도구분도 매핑 기법은 집계된 통계자료를 분해하여 세밀한 공간 단위로 나누거나, 행정구역과 같은 객체 단위의 데이터를 연속된 표면 모델로 만드는데 적용 가능한 기법으로 제안되고 있다. 본 연구에서는 밀도구분도 매핑의 유효성을 평가하기 위해 1) 연구대상지역의 사업체 정보를 행정동 단위의 통계로 집계하고, 2) 행정동 통계에 밀도구분도 매핑 기법을 적용하여 사업체의 분포를 구하고, 3) 공간 내삽에 의해 구해진 사업체의 분포와 원 정보로부터 구해진 분포를 비교하여 알고리즘의 성능을 평가하였다.

핵심어 : 밀도구분도 매핑, 공간 단위, 공간 내삽

## 1. 서 론

최근 GIS 연구의 주된 경향 중의 하나는 데이터의 통합과 융합이다. 원격탐사 영상과 인구통계자료와 같이 상호 연관되지 않는 것으로 여겨졌던 전통적인 데이터의 형식이 지금은 상호 보완적인 자료로 사용되고 있다. 예를 들어, 센서스와 지역지구 데이터가 순수 분광특성에 의한 기법의 한계를 뛰어넘을 수 있도록 자동화된 영상분류 의사결정 나무에 통합되어 사용된다(Hutchinson, 1982; Mesev, 1998). 이때, 인구통계의 기준이 되는 행정구역과 지역지

구 데이터의 경계는 서로 일치하지 않기 때문에 폴리곤 중첩기법을 이용하여 원 데이터의 공간 단위를 추가 분할하여야 하며 분할된 작은 공간 단위 폴리곤 각각에 어떻게 값을 부여하는가에 따라 분석 결과에 큰 영향을 주게 된다. 또한 원격탐사 영상과의 통합을 위해서는 벡터 형태의 폴리곤 데이터보다는 래스터 형태의 연속형 데이터가 필요하기 때문에 더 많은 공간 단위의 분할이 필요하다.

한편, 인구통계 데이터는 행정구역과 같은 공간 단위로 집계되어 제공되고 있다. 집계된 데이터를 활용할 때 발생하는 여러 가지 해석상의 문제는 대부분 공간 단위를

1) 정회원 · 한경대학교 토목공학과 겸임교수, 주식회사 타스텍 정보기술연구소장(E-mail:basil@naver.com)

분할하는데 개재되는 임의성에 의해 발생한다. 해석상의 문제 중 가장 현저히 나타나는 것은 '공간단위 수정가능성의 문제(Modifiable Areal Unit Problem)'로 알려져 있으며, 경계와 축척에 의한 집성의 정도가 공간 데이터의 분석 결과에 영향을 주는 현상으로 정의된다(Openshaw, 1983). 이 문제에 대한 잠재적인 해결 방법 중의 하나가 래스터 형태의 연속형 데이터로 인구통계를 표현하는 것으로서, 부정형으로 분할된 임의의 공간 단위에 종속되지 않고 연속된 장(field)으로 데이터를 모델링하는 것이다(Mennis, 2003).

이상에 살펴본 내용과 같은 현상은 특히 우리나라의 경우 통계청에서 집계 발간되는 인구통계의 최소 공간 단위가 행정동으로 일반적으로 사용되는 통계 구역에 비해 지나치게 크기 때문에 그 문제점이 더욱 두드러진다고 할 수 있다. 따라서 본 연구에서는 이러한 문제점을 극복하기 위해 제시되고 있는 여러 가지 방법 중 밀도구분도 매핑(Dasymetric Mapping) 기법이 실질적으로 적용될 수 있는지를 평가하는 것을 목적으로 하였다. 밀도구분도 매핑의 적용성을 평가하기 위해서는 모델링 결과의 정확도를 평가하여야 하며 인구통계 데이터의 경우 정확도 평가를 위한 모집단을 확보하기 어렵기 때문에 전화번호부의 사업체 정보를 원시자료로 활용하였다. 사업체 정보를 행정구역별로 집계하여 통계자료를 작성하고, 여기에 위성영상으로부터 추출된 토지피복정보와 도로망정보를 부가 데이터로 하여 밀도구분도 매핑을 적용하여 래스터 형태의 사업체수 추정 지도를 만들었다. 정확도 평가를 위해 원 사업체 정보에서 직접 내삽한 공간 단위별 밀도로부터 사업체수 분포도를 만들어 추정 결과와 비교하였다.

## 2. 연구동향

대부분의 인구통계 데이터는 연속형으로 제공되지 않고 있지만, 래스터 GIS의 분석 기법을 이용하여 센서스 정보로부터 인구와 인구특성을 연속형 데이터로 표현하기 위한 기법이 연구되고 있다. 이러한 연구 중 대표적인 것으로서 Flowerdew 등(1991)은 토지 이용별 공간 가중치를 부여하는 방법을 제안하였으며, Bracken 등(1989)은 공간 단위의 무게중심점을 이용한 내삽법을 제안하였고, Langford 등(1994)은 원격탐사 영상과 밀도구분도 매핑을 이용한 방법을 제안하였다(Mennis, 2003).

공간 내삽법으로 통칭할 수 있는 위의 기법 중 가장 단

순하면서 직접적인 방법은 공간 가중치를 부여하는 방법이다. 이 방법에서는 격자화소의 위치에 대응하는 공간 단위에 대한 각 격자화소의 면적 비에 따라 인구를 배분하는 방법이다.

Bracken 등(1989)은 영국의 통계구인 enumeration district의 인구통계를 연속형 데이터로 변환하기 위해 제한한 방법을 사용하였는데, 이 방법은 IDW(Inverse Distance Weighted) 내삽법을 변형한 것이다. 인구수를 공간 단위의 무게중심점과 같은 점사상에 부여하고, 이를 관측점으로 사용하여 거리제곱함수를 사용한 내삽법을 적용하여 래스터 데이터를 만들어 내었다. 주변에 위치한 중심점의 분포에 따라 이동 격자의 크기가 달라지는 이 방법에서는 래스터 데이터 중의 일부분이 인구수 0이 되는 것도 허용되었다.

래스터 형태의 인구통계 자료를 만들어 내기 위해 사용되는 또 하나의 기법이 밀도구분도 매핑에 의한 방법인데, 이는 공간 내삽을 지원할 수 있는 보조(부가 또는 관련된) 데이터를 사용하는 공간 내삽법이다. 이 기법은 임의의 공간 단위로 경계가 결정되지 않고, 내삽에 사용되는 변수 각각의 공간 분포가 반영된다는 점에서 공간 가중치 부여에 의한 방법과 차이가 있다. 이 기법은 처음에 인구통계의 인구를 지형도상의 거주가능지역과 거주불가능지역으로 구분하여 재배분한 것에서 시작되었다. 그 후 거주가능지역을 다시 주거유형 등에 따라 다시 구분하여 각각의 평균적인 인구밀도에 따라 인구수를 배정하였다. Langford 등(1991)은 Landsat TM 영상으로부터 추출된 토지이용도를 사용하여 토지이용별 인구밀도의 회귀모델을 구하고, 이를 기반으로 영국 Leicestershire의 인구통계 자료를 1km 크기의 래스터 데이터로 변환하였다. Mennis (2003)은 원격탐사를 이용하여 추출된 도시의 토지이용 데이터를 보조자료로 사용한 밀도구분도 매핑을 사용하여 미국 Pennsylvania 지역의 인구 센서스 자료로부터 100m 해상도의 연속형 인구 자료를 생성하였다.

## 3. 밀도구분도 매핑 기법

밀도구분도 매핑 기법은 통계 구역에 할당된 계수(실측값) 데이터와 보조 데이터를 입력 데이터로 사용하여 통계 구역과 보조 데이터의 클래스별 구역이 교차되는 출력 구역으로 데이터를 재배분한다. 데이터의 재배분은 공간 가중치와 보조 데이터 클래스의 상대적 밀도를 기준으로 이루어진다. 통계 구역을  $s$ , 보조 데이터의 클래스  $c$ 에

대응되는 구역을  $z$ 라고 하면 출력구역  $t$ 는  $s$ 와  $z$ 가 중첩되는 영역으로 정의된다. 주어진 출력구역에 대한 계수의 추정치는 다음 식 (1)과 같이 계산된다.

$$\hat{y}_t = y_s \left( \frac{A_t \hat{D}_c}{\sum_{t \in s} (A_t \hat{D}_c)} \right) \quad (1)$$

여기서  $\hat{D}_c$ 는 보조 데이터 항목  $c$ 의 추정 밀도,  $y_s$ 는 통계 구역의 계수,  $\hat{y}_t$ 는 출력 구역의 추정 계수,  $A_t$ 는 출력 구역의 면적이다.

$\hat{D}_c$ 는 분석자가 해당 클래스에 대한 밀도값을 선형적으로 알고 있다면, 사용자가 지정할 수 있으며, 그렇지 않은 경우 통계 구역들로부터 보조 데이터의 클래스에 대응되는 구역을 표본 추출하여 계산할 수도 있다. 표본 추출의 방법은 세가지가 있다. ‘포함’ 조건은 통계 구역이 완전히 보조 데이터의 한 클래스에 포함되는 경우이다. ‘중심점’ 조건은 통계 구역의 무게중심점이 보조 데이터의 한 클래스에 포함되는 경우이다. ‘중복도’ 조건은 통계 구역이 어느 클래스에 포함되는 것으로 할지를 중복도에 따라 결정하는 경우이다. 이러한 표본추출은 통계 구역과 출력구역의 크기가 유사할 경우에 적용이 용이하며, 폴리곤인 통계구역을 훨씬 작은 단위의 래스터형 출력구역으로 변화하는 경우에는 적용이 곤란한 단점이 있다. 표본 추출된 경우에  $\hat{D}_c$ 는 식 (2)를 이용하여 계산된다.

$$\hat{D}_c = \frac{\sum_{s=1}^m y_s}{\sum_{s=1}^m A_s} \quad (2)$$

여기서  $m$ 은 보조 데이터의 클래스  $c$ 에 대응되는 통계 구역으로부터 추출된 표본의 크기,  $A_s$ 는 통계 구역의 면적이다.

이 때, 보조 데이터의 클래스 중 일부에 대해서라도 분석자가 계수가 있을 수 없는 것을 알 수 있는 클래스가 있다. 예를 들어 인구에 대해 토지이용이 수계인 경우에는 인구수가 할당되지 않을 수 있다. 이와 같이  $\hat{D}_c$ 는 계수가 존재할 수 있는 보조 데이터의 클래스에 대해서만 정의된다.

데이터의 밀도와 보조 데이터 클래스 간의 상관관계의 공간적 변동을 고려하기 위해 밀도구분도 매핑 기법에서는 추가적인 보조 데이터를 더 사용할 수 있다.

실제 GIS에서는 벡터의 중첩분석이나 래스터의 중첩분석을 모두 이용할 수 있지만 Hultgren(2004)의 연구에 따

르면 래스터의 중첩분석이 작업시간이 월등히 짧은 장점이 있다.

## 4. 모의실험 및 평가

대부분의 기존 연구에서는 인구통계를 대상으로 모델링이 적용되었으나, 본 연구에서는 알고리즘의 효율성을 평가하기 위한 정확도 검증 작업이 필요하였기 때문에 정확도 검증이 가능한 통계를 모델링 대상으로 사용하고자 하였다. 개인이 모집단이 되는 인구통계의 경우 정확도 검증을 위한 기준자료의 작성 또는 확보가 불가능하기 때문에 많은 선행연구에서 검증된 안정성에도 불구하고 사용할 수 없었으며, 그에 대신하여 모집단을 확보할 수 있는 상호 데이터를 사용하였다.

본 연구에서 분석을 위해 사용된 사업체 데이터는 2003년 2월 기준 전화번호부의 서초, 강남 업종편 정보로부터 추출한 84,683개의 상호이며, 데이터의 처리에는 MS Access와 범용 GIS 소프트웨어인 ArcGIS 8.3과 ArcInfo 8.1.2를 이용하였고, 소프트웨어에서 지원되지 않는 주소정제 및 주소매칭은 Visual Basic을 이용하여 프로그래밍하였다.

또한 보조 데이터로는 위성영상을 분석하여 7개의 클래스로 분류된 토지피복도와 주요 도로변에 밀집되는 상업 및 업무시설의 특성을 감안하여 주요 간선 도로망을 사용하였다.

### 4.1 통계자료 작성

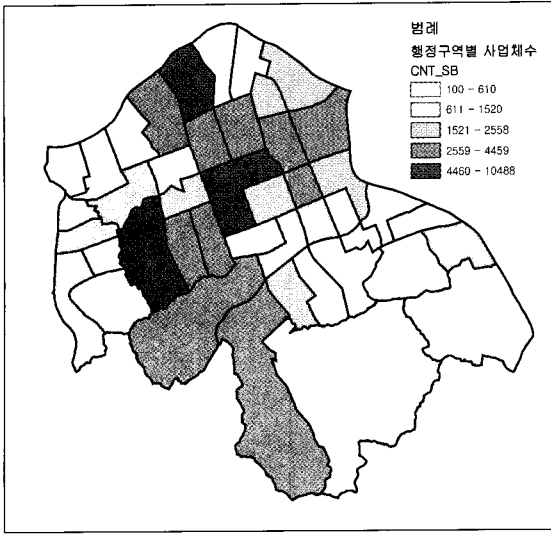
그림 1과 같이 점사상으로 표현된 상호 데이터에 행정 구역 레이어를 중첩하여 행정구역별로 상호 데이터를 집계하여 행정구역별 사업체수를 집계하고, 행정구역의 면적으로 나누어 행정구역별 사업체밀도를 산출하였다.

### 4.2 기준자료 작성

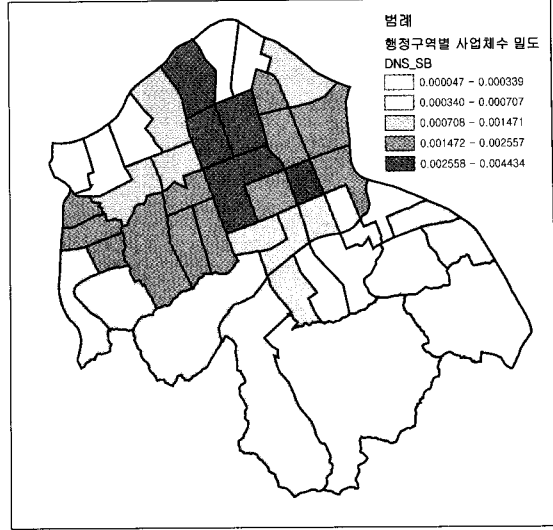
통계작성을 위해 사용된 상호로부터 반경 100m 내의 사업체 밀도를 구한 후, 전체 사업체수를 밀도의 합으로 나눈 값을 구해진 밀도에 곱하여 사업체의 분포도를 작성하고(그림 2), 이를 정확도 평가를 위한 기준자료로 사용하였다.

### 4.3 보조 데이터

밀도구분도 매핑을 위한 보조 데이터로 토지피복도와 도로망도를 사용하였다(그림 3). 토지피복도는 환경지리정

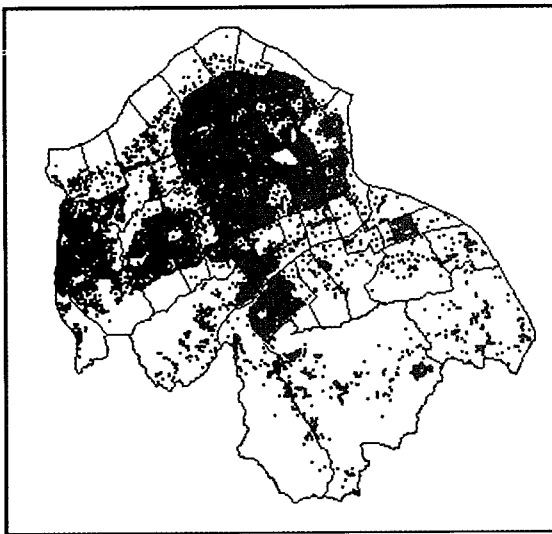


(a) 행정구역별 사업체수

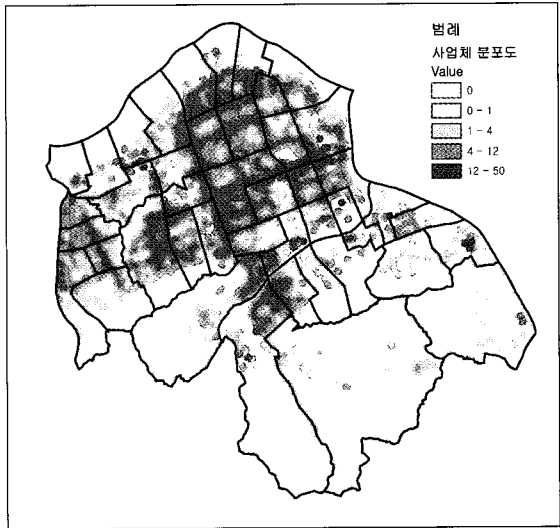


(b) 행정구역별 사업체 밀도

그림 1. 통계자료 작성



(a) 상호 데이터



(b) 사업체 분포도

그림 2. 기준자료 작성

보(egis.mc.go.kr)에서 구축한 토지피복분류도를 사용하였으며, 도로망도는 수치지도로부터 세도로를 제외한 나머지 도로를 통합하여 사용하였다. 또한 도로로부터 100m의 버퍼를 적용하여 도로 인접지역을 구하여 활용하였다.

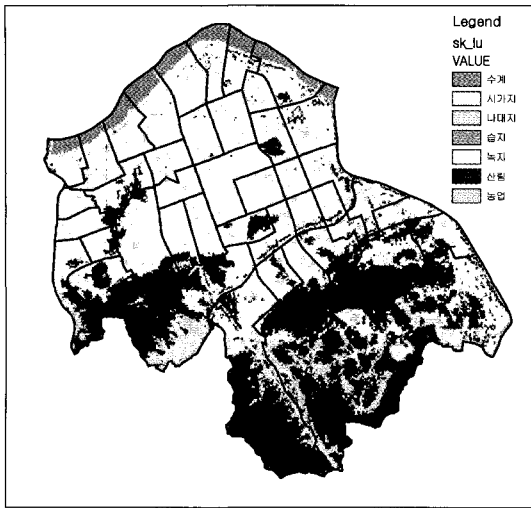
#### 4.4 밀도구분도 매핑

본 연구에서 적용한 밀도구분도 매핑 알고리즘은 그림

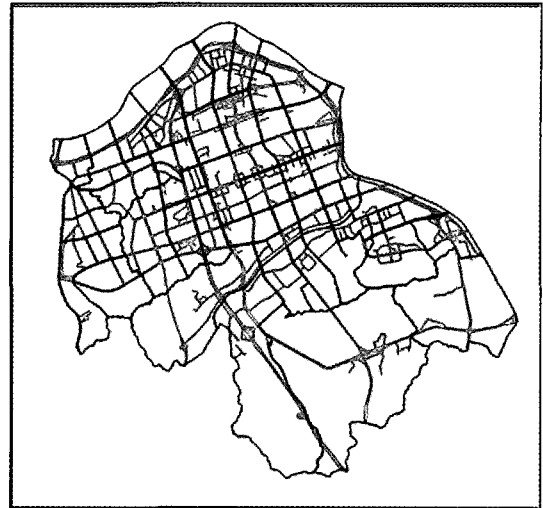
4와 같은 순서로 이루어 졌다.

순서도의 각 단계를 상세히 설명하면 다음과 같다.

1. 행정동별 사업체수 밀도와 도로망도를 토지피복도 (30m 격자)에 맞추어 래스터 데이터로 변환한다.
2. 행정동별 사업체수 밀도에 격자의 면적을 곱하여 격자별 사업체수를 구한다.
3. 도로로부터 100m 버퍼 영역을 구한다.



(a) 토지피복도(자료제공 : 환경부)



(b) 도로망도

그림 3. 보조 데이터

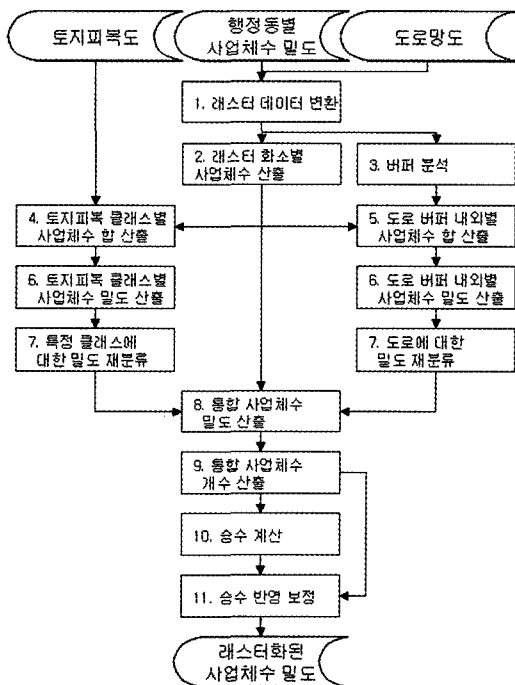


그림 4. 연구에 사용된 밀도구분도 매핑 알고리즘 순서도

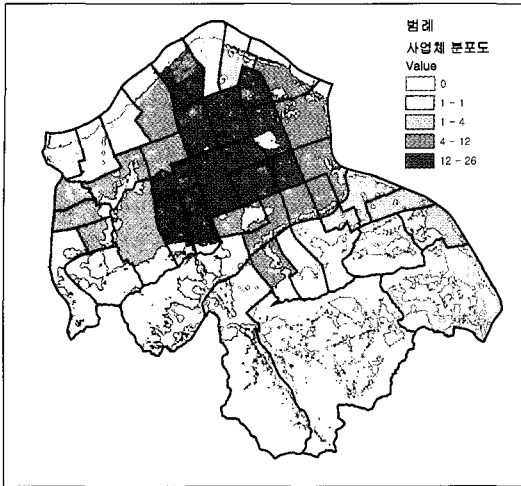
4. 토지피복 클래스별 면적과 중복된 행정동 격자별 사업체수의 합을 구한다.
5. 도로와 도로 버퍼 별 면적과 중복된 행정동 격자별 사업체수의 합을 구한다.

6. 토지피복 클래스별, 도로/도로버퍼별 사업체수 밀도를 구한다.
7. 수계, 산림, 농지와 같은 특정 토지피복 클래스 및 도로에 대한 사업체수 밀도를 0으로 재분류(reclassify) 한다.
8. 행정동별 사업체밀도, 토지피복 클래스별 사업체밀도, 도로/도로버퍼별 사업체수 밀도를 곱하여 격자별 사업체 밀도를 구한다.
9. 추정된 사업체 밀도에 격자의 면적을 곱하여 격자별 추정 사업체수를 구한다.
10. 행정동별로 격자별 추정 사업체수를 각각 더하고, 행정동별 원 사업체수와 비교하여 행정동별 승수를 구한다.
11. 9.에서 구해진 추정 사업체수에 10.에서 구해진 승수를 곱하여 최종 추정 사업체수를 구한다.

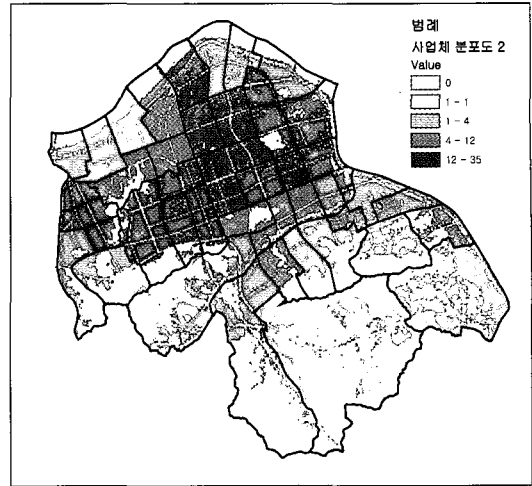
#### 4.5 결과 및 평가

본 연구에서는 보조 데이터의 추가에 따른 효과를 파악하기 위해 토지피복도만 사용하여 모델링된 결과와 도로망까지 사용하여 모델링된 결과를 모두 구하였다. 결과의 평가를 위해서 기준자료인 그림 2(b)의 사업체 분포도와 의 평균제곱근오차(RMSE)와 상관계수를 분석하였다.

행정동별 사업체수 통계자료에 밀도구분도 매핑 기법을 적용하여 산출된 결과는 그림 5와 같다. 시각적 평가를 위하여 강남역 인근지역을 확대하여 제시하였다(그림 6). 그림 5와 그림 6에서도 알 수 있는 바와 같이 밀도구분도 매핑 기법을 이용하여 통계자료를 래스터화한 결과

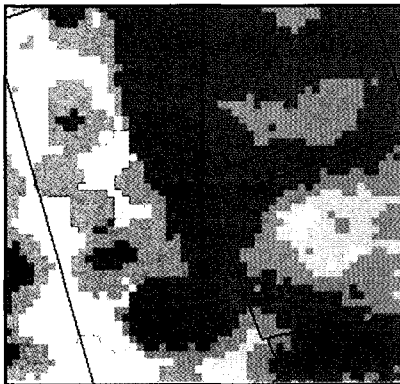


(a) 토지피복도만 사용한 경우

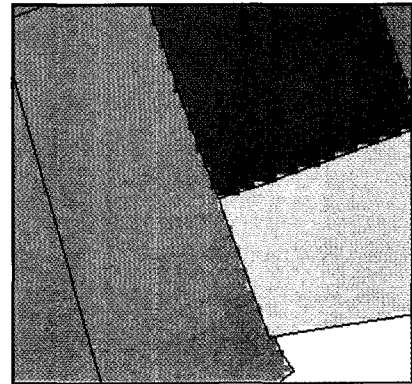


(b) 도로망도까지 사용한 경우

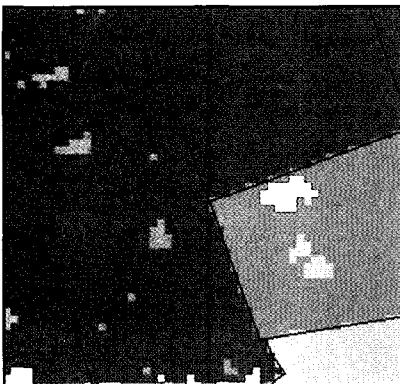
그림 5. 밀도구분도 매핑에 의한 사업체수 분포도



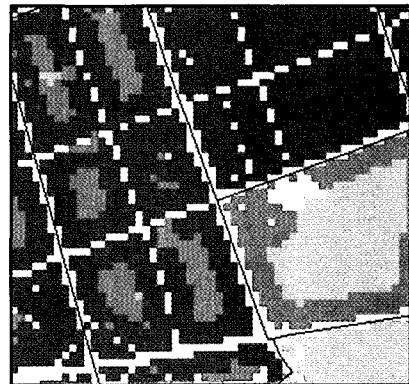
(a) 기준 데이터로 생성된 사업체수 분포도



(b) 행정구역별 사업체수 분포도



(c) 토지피복도만 사용한 사업체수 분포도



(d) 도로망도까지 사용한 사업체수 분포도

그림 6. 사업체수 분포도 (강남역 인근 확대)

표 1. 추정 사업체 분포도 품질 평가 결과

항목	평균제곱근오차	상관계수
기준 데이터 vs. 행정동별 사업체수	8.76	0.53
기준 데이터 vs. 토지피복만 사용한 사업체수	6.95	0.60
기준 데이터 vs. 도로망까지 사용한 사업체수	6.62	0.63

는 시각적으로는 상당한 자료 품질의 향상을 확인할 수 있었다. 그림 6의 (b)와 (c), (d)를 비교하면 행정구역 단위로 하나의 값만을 가지던 분포도에서 단일 행정구역 내에서도 위치에 따라 서로 다른 값을 갖는 분포도를 갖는 것을 볼 수 있다.

시각적인 자료 품질의 향상이 실질적으로 이루어진 것 인지를 평가하기 위해 기준 데이터에 대해 표 1과 같이 평균제곱근오차와 자료의 상관성을 분석하였다. 상관성은 같은 위치에 있는 래스터 데이터 값 간의 상관관계를 구하는 Arc/Info Grid 모듈의 correlation 명령어를 사용하여 평가하였으며, 평균제곱근오차는 식 (3)을 사용하여 평가하였다.

$$\sqrt{\frac{\sum(\hat{y}-y)^2}{n}} \quad (3)$$

여기서  $\hat{y}$ 는 구해진 추정 사업체수,  $y$ 는 기준 데이터의 사업체수,  $n$ 은 화소수 이다.

본 연구에서 사용된 부가 데이터에 대해서는 연구자가 클래스별 밀도에 대한 선행적 지식을 가지고 있지 못했을 뿐더러, 도로망 데이터에 대해서도 한쪽에서는 아파트 단지내 도로가 들어가 있는데, 다른 쪽에서는 주요한 이면도로도 빠져 있는 등 데이터 품질의 안정성이 확보되지 않아 추가적인 좋은 결과를 기대하기는 어려웠지만, 표 1에 제시되어 있는 바와 같이 부가 데이터를 더 넣음에 따라 추정된 통계와 원 데이터 간의 오차도 줄어들고 분포되는 양상도 유사해지는 것으로 나타나 밀도구분도 매핑에 의한 통계 데이터의 래스터화 기법의 활용 가능성은 충분히 있는 것으로 평가된다.

또한, 토지피복도보다는 토지이용현황도와 같은 인간 활동에 보다 직접적으로 연관된 데이터를 보조 데이터로 활용하면 더욱 향상된 분석결과를 얻을 수 있을 것으로 기대된다.

## 5. 결 론

본 연구에서는 GIS 및 원격탐사 분야에서 다양한 통계 데이터를 통합 사용하기 위해 해결해야할 공간단위 수정가능성의 문제와 공간 단위 설정의 문제에 대한 적절한 해결책의 하나로 제시되고 있는 밀도구분도 매핑 기법에 대해 고찰하고 모의실험을 통해 그 효용성을 평가하였다.

본 연구의 실험에서 밀도구분도 매핑 기법을 적용하여 통계데이터를 래스터화 한 결과 통계데이터를 단순히 래스터화 한 것에 비해 평균제곱근오차가 8.76에서 6.62로 상관계수가 0.53에서 0.63으로 각각 개선되어 밀도구분도 매핑 알고리즘의 적용을 통해 데이터의 밀도 및 설명력을 높일 수 있음을 확인할 수 있었다. 또한, 토지피복만을 부가 데이터로 사용한 결과에 비해 도로망도까지 부가 데이터로 사용한 결과가 평균제곱근오차는 작아지고 상관계수는 커지는 것으로 볼 때 적절한 부가 데이터의 추가에 의해 밀도구분도 매핑이 더욱 정교해 질 수 있음을 알 수 있었다.

밀도구분도 매핑 기법은 알고리즘이 비교적 단순하면서도 통계 데이터와 같은 집성된 폴리곤 데이터로부터 세밀한 래스터 데이터를 직접 만들어 낼 수 있는 기법이기 때문에 인구통계와 같은 정보를 원격탐사 영상분석에 적용할 때, 매우 유용한 기술로 평가된다.

이와 같이 밀도구분도 매핑 기법은 제한된 통계자료를 이용하여 복잡한 분석을 수행해야하는 분석자나 연구자에게 매우 유용한 기법이나, 적절한 부가 데이터의 선택 및 부가 데이터 클래스에 대한 선행적 지식 유무가 결과에 상당한 영향을 주기 때문에 이에 대한 프레임웍을 제시해 줄 수 있는 추가적인 연구 또한 필요할 것으로 판단된다.

## 참고문헌

- Bracken I. and Martin D. (1989), The generation of spatial population distributions from census centroid data, *Environment and Planning*, Vol. 21, pp. 537-543. in Mennis J. (2003), Generating surface models of population using

- dasymetric mapping, *The Professional Geographer*, Vol. 55, No. 1, pp. 31-42.
- Flowerdew R. and Green M. (1992), Developments in areal interpolation methods and GIS, *Annals of Regional Science*, Vol. 26, pp. 67-78. in Mennis J. (2003), Generating surface models of population using dasymetric mapping, *The Professional Geographer*, Vol. 55, No. 1, pp. 31-42.
- Hultgren T. (2004), Raster-based automated dasymetric mapping, *UCGIS 2004*, <http://www.ucgis.org/ucgisfall2004/studentpapers/files/hultgren.pdf>
- Hutchinson C.F. (1982), Techniques for combining Landsat and ancillary data for digital classification improvement, *Photogrammetric Engineering and Remote Sensing*, Vol. 48, pp. 123-130.
- Langford M., Maguire D., and Unwin D. (1991), The areal interpolation problem: Estimating population using remote sensing in a GIS framework, *Handling Geographical Information: Methodology and Potential Applications*, ed I, Masser and M. Blakemore, pp. 55-77., Longman, London
- Langford M. and Unwin D. (1994), Generating and mapping population density surfaces within a geographic information system, *The Cartographic Journal*, Vol. 31, pp. 21-26. in Mennis J. (2003), Generating surface models of population using dasymetric mapping, *The Professional Geographer*, Vol. 55, No. 1, pp. 31-42.
- Mennis J. (2003), Generating surface models of population using dasymetric mapping, *The Professional Geographer*, Vol. 55, No. 1, pp. 31-42.
- Mesev V. (1998), The use of census data in urban image classification, *Photogrammetric Engineering and Remote Sensing*, Vol. 55, pp. 31-42.
- Openshaw S. (1983), *The modifiable areal unit problem*, Concepts and Techniques in Modern Geography, 38. Geo Books, Norwich, UK.

---

(접수일 2006. 10. 11, 심사일 2006. 10. 30, 심사완료일 2006. 10. 30)