STATISTICS KOREA

Progress and Future Plans for Big Data Use at Statistics Korea

Statistics Korea

Workshop on the modernization of statistical production

15~17 Apr 2015, Geneva

STATISTICS KOREA

Contents

Background of big data getting focus

- **Government policy to foster big data usage**
  - Korean government established a 'big data master plan' at national science and technology commission on Nov, 2012
  - Also established national infrastructure for big data sharing, provision of technical support and expert training
  - Statistical production using big data composes a part of the big data master plan
- **Request to improve statistics to fit the needs of people**
  - The actual economic and social sentiment of people is not reflected at national statistics
  - Statistical production process should be improved to produce more temporal and practical statistics
- **Statistical production environment has changed**
  - Increased data collection costs and response burden
  - Evolution in data collection paradigm: survey -> administrative data -> big data
  - Need more modernized and effective methods of statistical production

Progress on big data application by Statistics Korea

- **Mining and Manufacturing Production Index Support(2012)**
  - Support rapid and accurate production of the mining and manufacturing production index by using media data
  - Media data is automatically collected and used for objective editing(checking changes in establishments and items, finding outliers) before finalizing the index
  - Easier understanding of time-series data by providing visualized analysis function
- **Online Price Index Reporting System(2013)**
  - Collect and standardize online price information from websites, calculate and visualize online price index
  - Used to improve accuracy of the official statistics by recognizing price fluctuation of the major price index items promptly and conducting comparative analysis with official CPI
- **Daily Population Movement Analysis System(2014)**
  - Construct daily population movement data using GPS information of the mobile users
  - Implement GIS-based analysis system to support policy decision by the various organizations
- **Research on Statistical Production System using Big Data(2014)**
  - Analyze major issues regarding big data use on the national statistical production and propose a future plan

Data Construction

- Purchase GPS location data from a domestic mobile carrier(SK Telecommunication)
  - Estimate the total number of population movement by reflecting the market share
- To deal with privacy issue, data is aggregated as a certain unit before provided to Statistics Korea

Monthly Information

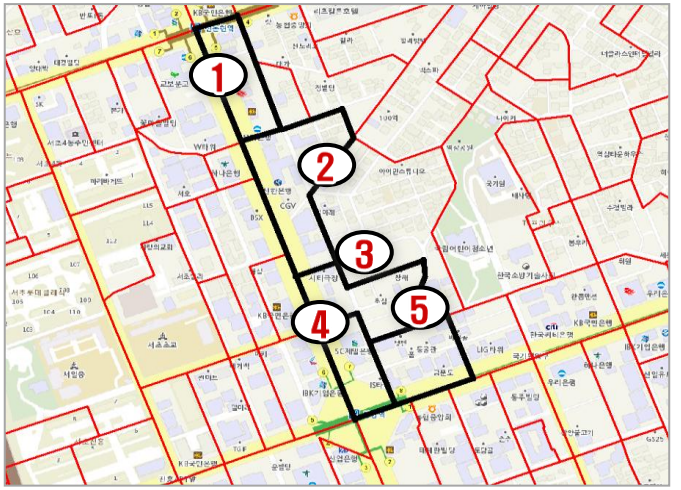
- 50m*50m unit information, consist of sex/age of the mobile users and day/time information.
- Offer daily average population

50m Unit monthly info

Daily Information

- Data aggregated by the area pre-offered by Statistics Korea, consist of sex/age/time and origin/destination
- Offer number of people visiting the area within the given time period

Daily info aggregated by 'aggregated area'



Data Construction

- Obtain the number of visitors by excluding the number of resident
  - Set the actual residence field using the night-time location data

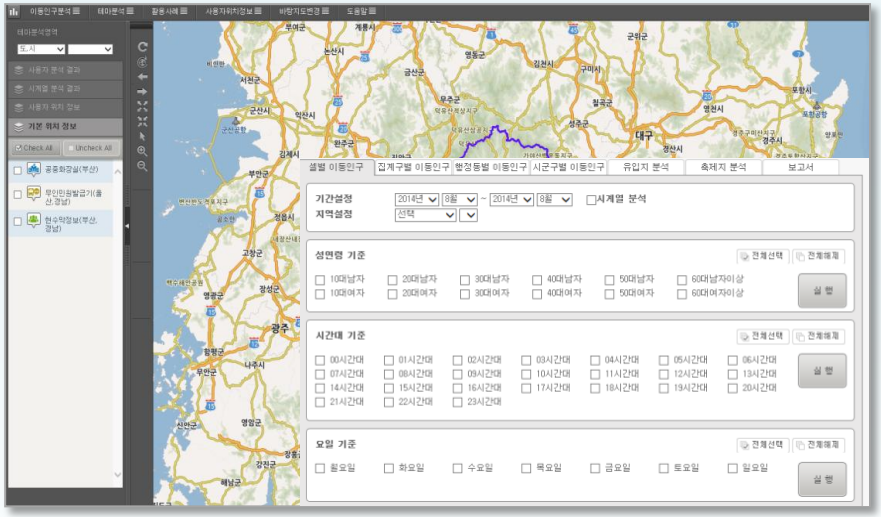
  (Actual residence is assigned as a place where the mobile user mostly resides on weekday night between 1 am to 5 am for 3 months)

  - Extract the number of visitors using the actual residence information
- Compensate sex/age/market share deviation
  - For the children of age 0 to 9 who don't have mobile phones, conduct residence-based compensation
  - Compensate regional market share deviation using the number of floating population(directly measured at 13,000 branches)

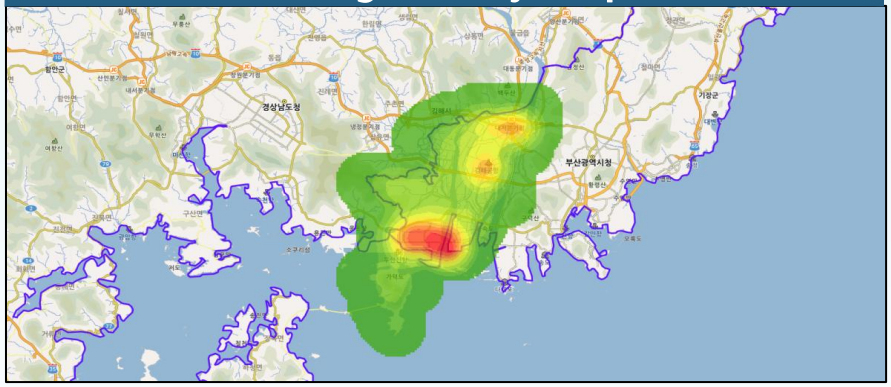
Primary functions

Function	Description
Population movement analysis	• Analyze and visualize population movement by 50x50 cells/aggregated areas/administrative districts/origins • For the pre-selected festivals, analyze and visualize the number and the features(sex/age) of the festival visitors
Theme analysis	• Analyze and visualize quarterly/seasonal/commuting/weekend/weekdays/hourly population movement
Use case	• Provide use case scenario of the system and the sample data to test on the map
User information analysis	• Upload personal user xsl file(location-based) • Provide custom analysis and visualization function
Other features	• Switch background map between the vector map and the satellite image

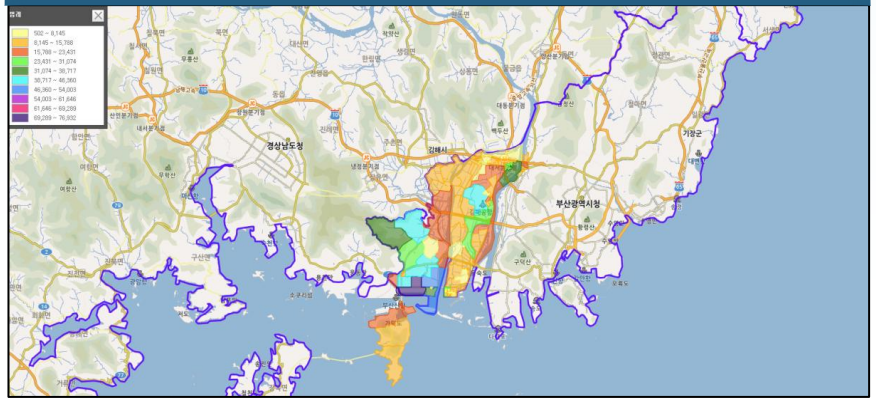
Visualization example



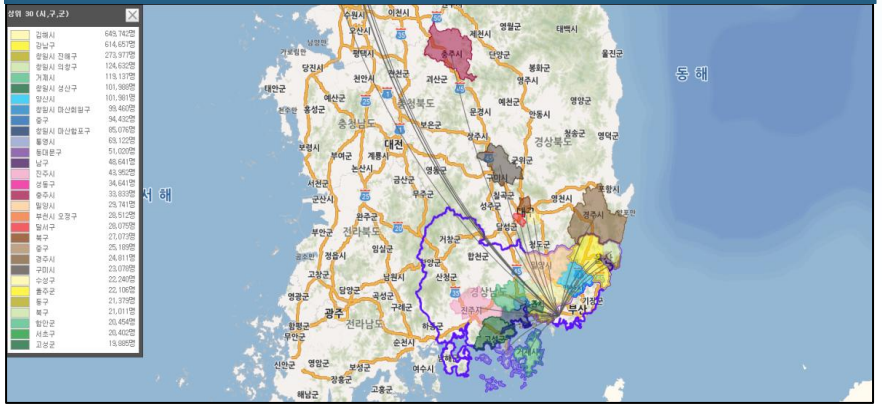
Display regional population movement using density map



Theme map



Festival analysis

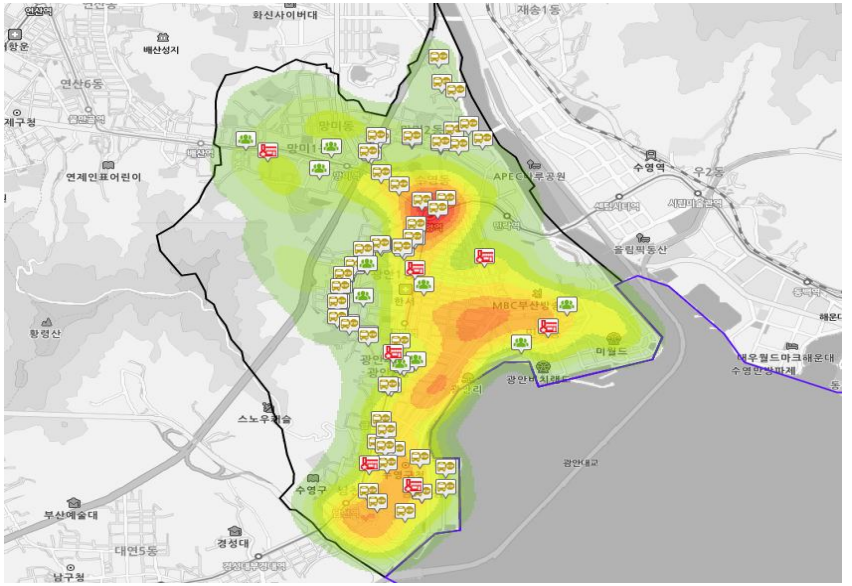


Use case 1 – midnight local bus schedule adjustment

Description

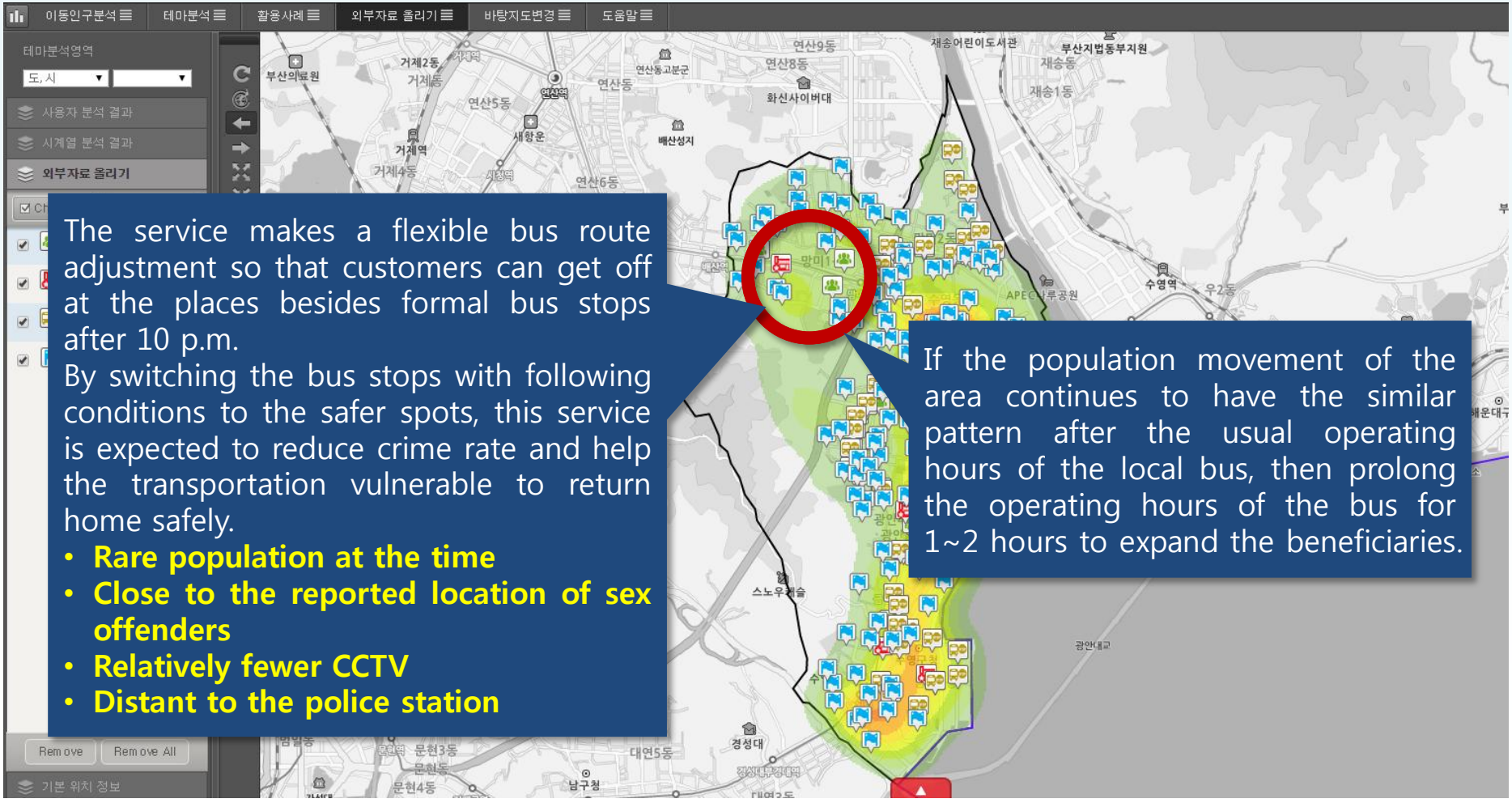
Purpose	• Flexible adjustment on the routes and operating hours of the local bus to help women and teenagers to return home safely • Make a stopover to avoid the lonesome and dangerous places at midnight
Data to be used	• Location information of the sex offenders • Midnight population movement data • Location information of the police stations and patrol divisions

Benefits



Benefits	• Flexible and efficient operation of local buses at midnight • Crime deterrent effect

Use case 1 – midnight local bus schedule adjustment



The service makes a flexible bus route adjustment so that customers can get off at the places besides formal bus stops after 10 p.m.

By switching the bus stops with following conditions to the safer spots, this service is expected to reduce crime rate and help the transportation vulnerable to return home safely.

- **Rare population at the time**
- **Close to the reported location of sex offenders**
- **Relatively fewer CCTV**
- **Distant to the police station**

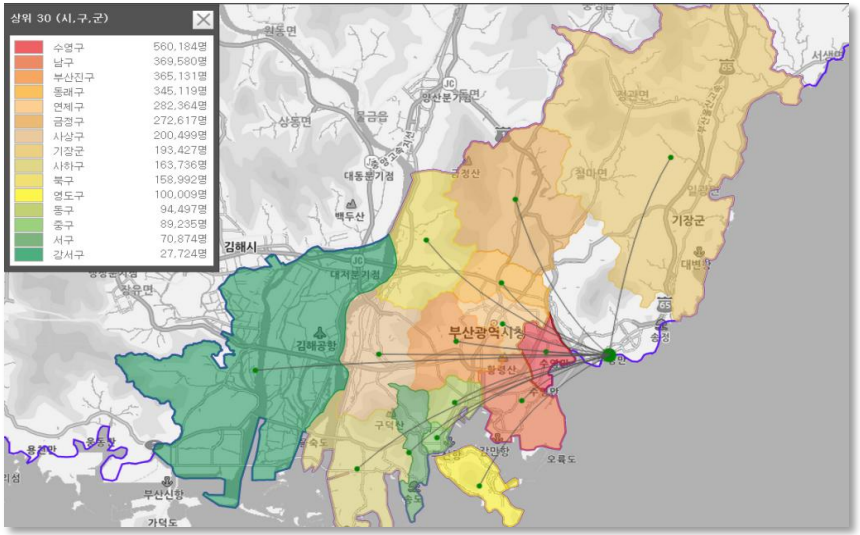
If the population movement of the area continues to have the similar pattern after the usual operating hours of the local bus, then prolong the operating hours of the bus for 1~2 hours to expand the beneficiaries.

Use case 2 – shuttle bus service at festival place

Description

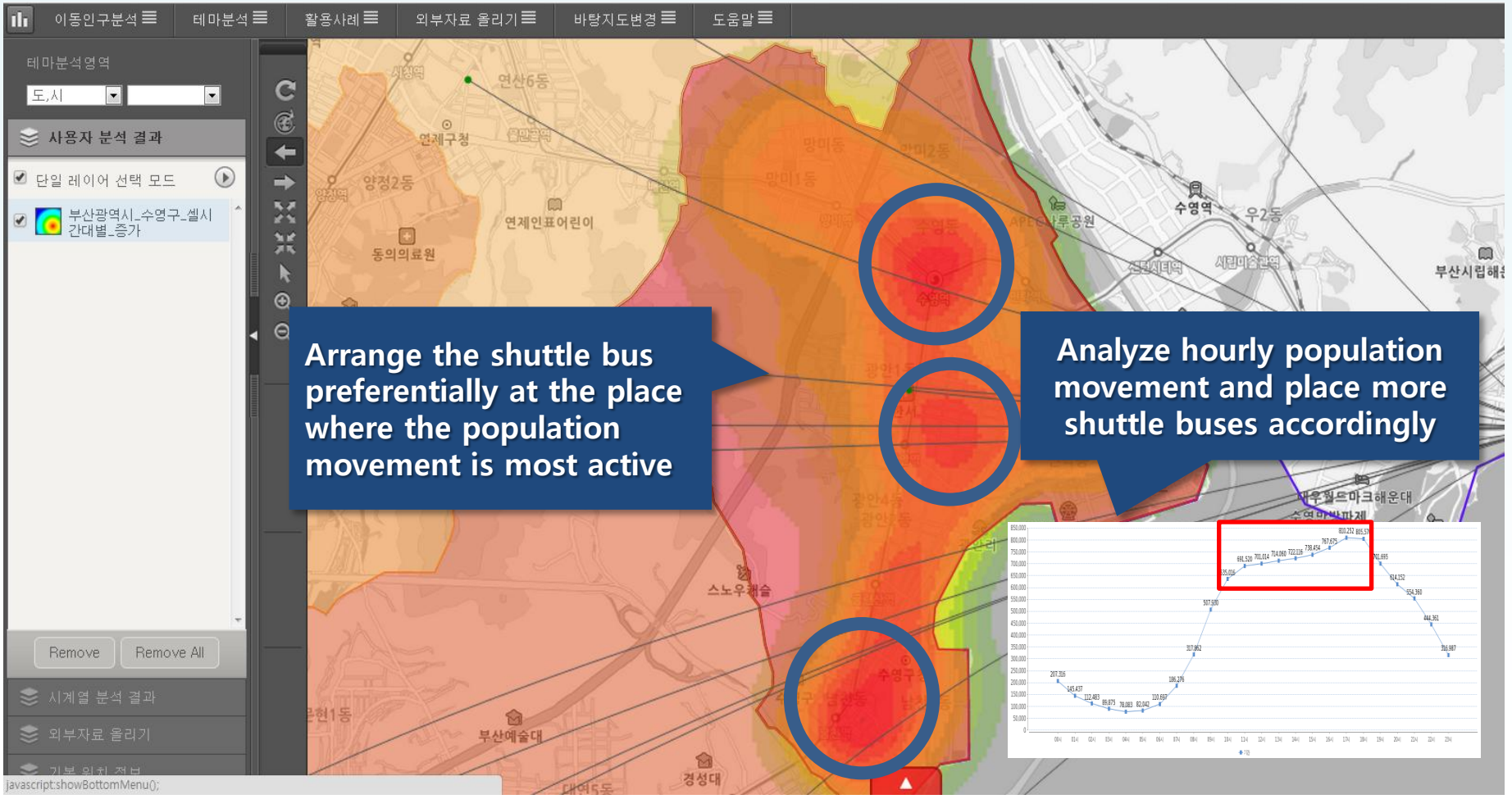
Purpose	• Using festival population analysis, effectively schedule shuttle bus service in the festival period
Data to be used	• Population influx data between the festival place and the neighboring areas • Population movement data between the festival place and the surrounding tour spots

Benefits



Benefits	• Increased convenience to access transportation • Relieve traffic congestion and secure parking space in the festival period

Use case 2 – shuttle bus service at festival place



Use case 3 – holiday population movement analysis

합계 : 방문객(명)	열 레이블																
행 레이블	강서구	금정구	기장군	남구	동구	동래구	부산진구	북구	사상구	사하구	서구	수영구	연제구	영도구	중구	해운대구	총합계
부산광역시	1054250	2843209	1548835	3066952	1421168	3555911	5242615	2824739	3222367	3341060	1430141	2208836	2551086	1418310	1248971	5084762	42063212
경상남도	280203	223409	114060	152298	67696	214322	324198	253338	279552	185799	90328	107930	119914	70590	81368	226950	2791955
김해시	125739	52087	28158	50694	28668	75068	124817	114901	116635	65388	30171	36309	42864	25290	27881	79319	1023989
양산시	20306	130193	70036	33531	16530	66904	67473	62558	39443	25122	13449	27450	36495	12422	14725	74450	711087
거제시	22967	15738	8213	19092	10021	17447	37997	15176	23146	24447	13161	12996	12766	13998	12813	28423	288401
창원시 성산구	16846	13736	7653	15316	5978	16467	26796	12985	19066	11917	6603	10953	10495	4924	5976	23932	209643
창원시 진해구	47100			19063	6499		23657	11428	20011	27109	8589	10336	8331	8351	8412		198886
창원시 의창구	16429	11655		14602		14433	24322	13348	19187	13224	7012	9886	8963	5605	6282	20826	185774
진주시	8998					11879	19136	11268	15427	9242	5540						81490
창원시 마산회원구	11703					12124		11674	15451	9350							60302
창원시 마산합포구	10115								11186								21301
통영시											5803				5279		11082
충청북도	205081	48104	69096	56364	40186	56928	95331	56587	197324	165886	30011	40389	60340	26498	30577	107782	1286484
청원군	205081	48104	6														84
경기도	20827	63104	1														92
성남시 분당구	13567	63104	1														82
부천시 오정구	7260																60
울산광역시	11234	96163	14														76
남구	11234	35819	5														61
울주군		19604	4														54
중구		16432															77
동구		12990															86
북구		11318															98
(비어 있음)	19261	49364	2														43
	19261	49364	2														43
서울특별시	8956	17113	12828	38451	23689	32961	50743	22625	32667	29260	13740	47238	21833	12269	13950	124206	502529
강남구	8956	17113	12828	18707	10010	19343	29393	12261	18299	17107	8180	18726	12979	6818	8257	47679	266656
중구				19744	6773	13618	21350	10364	14368	12153	5560	8950	8854	5451	5693	21565	154443
서초구												10634				31419	42053
송파구												8928				23543	32471
마포구					6906												6906
경상북도		12117	15322		5953					10759							44151
구미시			6782		5953					10759							23494
경주시		12117	8540														20657
총합계	1599812	3352583	1955476	3515236	1663017	4075108	6044395	3280102	3875825	3845422	1643309	2570216	2932142	1594653	1458684	5878462	49284442

This display shows the population movement analysis in holiday period. The result shows the origin of the population movement to the selected city inside the given scope.
Using the analysis result above, administrators can improve transportation accessibility for the home visitors by flexible adjustment of the express/ intercity bus and train schedules.

Implication of the past projects

- **Publishing official statistics using big data is a difficult issue**
  - Big data usually have a strong sampling bias so that quality control is considerably hard
  - Publishing official statistics using big data is our ultimate goal, but for now the usage of big data is limited to administrative support
- **Need to establish a long-term roadmap to consolidate unstructured data to structured statistical production system**

6 Challenges on the big data use at statistical production system

Legal issue

- There's no legal basis to obtain big data from the data-owning agencies under the present statistics law
- Need institutional strategy to share big data with governmental/private agencies

Cost effectiveness

- Cost incurs when building IT infrastructure to collect/ process big data and purchasing data
- Need detailed consideration on cost performance

Privacy issue

- Need legal mechanism to balance between privacy and statistical use
- Need to maintain public trust that private information is strictly protected and used only for statistical purpose

Methodology

- According to the awareness investigation, methodological issue is as important as legal and privacy issue
- Need new analyzing method such as data mining to handle big data
- Educate existing employees and hire experts from outside

Data management

- Importance of data security is gaining notice since the increase of administrative data use
- Data management method should be customized at each statistics published using big data

Information Technology

- statistical production technology using big data would be drawn from the interface between the existing statistical technology and the general big data technique
- Needed to define big data techniques to adopt at each steps of data processing

Statistical production process should be revised

- Statistics Korea has been following own standard KSBPM(revised version of GSBPM)

1. Plan	2. Design	3. Build	4. Collect	5. Process	6. Analyze	7. Disseminate	8. Archive	9. Evaluate
1.1 Determine statistical demand	2.1 Design output	3.1 Build collection instrument	4.1 Select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Prepare dissemination data	8.1 Define archive rules	9.1 Make evaluation plan
1.2 Verify & Specify statistical demand	2.2 Design variables	3.2 Build production system	4.2 Prepare collection	5.2 Classify & code	6.2 Validate outputs	7.2Produce disseminate products	8.2 Manage archive repository	9.2 Conduct evaluation & produce reports
1.3 Establish output objectives	2.3 Design collection methodology	3.3 Configure workflows	4.3 Run collection	5.3 Review, validate & edit	6.3 Scrutinize & explain	7.3 Manage release of dissemination products	8.3 Preserve data & associated metadata	9.3 Derive improvement plans & make action plan
1.4 Identify statistical concepts	2.4 Design universe & sample	3.4 Test production system	4.4 Finalize collection	5.4 Impute	6.4 Apply disclosure control	7.4 Promote dissemination Products	8.4 Dispose of data & associated metadata	
1.5 Check data availability	2.5 Design processing methodology	3.5 Test business process		5.5 Derive new variables & statistical units	6.5 Finalize outputs	7.5 Manage user support		
1.6 Make production plan	2.6 Design production system	3.6 Finalize production system		5.6 Calculate weights				
				5.7 Calculate aggregates				
				5.8 Finalize data processing				

※ KSBPM : 9 phases and 47 processes

Statistical production process should be revised

- Based on GSBPM v5.0, revise the process reflecting the characteristics of big data

Plan	Design	Build	Collect/ Process	Analyze	Disseminate	Evaluate
Determine & verify statistical demand	Consult with the agencies concerned	Build collection instrument	Collect data and review	Prepare draft outputs	Prepare dissemination data	Make evaluation plan
Establish output objectives	Analyze raw data	Build production system	Integrate data	Validate outputs	Produce disseminate products	Conduct evaluation & produce reports
Identify statistical concepts	Design output	Build Dissemination system	Classify & code	Scrutinize & explain	Manage release of dissemination products	Derive improvement plans & make action plan
Check data availability	Design variables	Configure workflows	Impute	Visualize	Promote dissemination Products	
Make production plan	Design collection methodology	Test production system	Derive new variables & statistical units	Apply disclosure control	Manage user support	
	Design processing methodology	Test system security	Selection (Filtering from raw DB)	Finalize outputs		
	Design production system	Test business process	Calculate weights			
		Finalize production system	Calculate aggregates	Finalize data processing		

Newly added sub-process

Statistical methodology needed to process big data

- **Example methodologies need to be reviewed to produce indicators using big data**
  - Besides the ones described below, additional statistical modeling technology(prediction, classification, clustering, …) should be reviewed for future use

		Description
Search duplicated /missing values		• Verify whether records are duplicated/missing promptly after data collection • Classify data, add items and verify the records are classified and matched appropriately • Analyzing tools(OLAP, …) are used to do the job
Imputation		• Refine data by replacing missing/delayed data with the imputed value • Process automatically by Rule-based programming
Statistical modeling	ARIMAX	• For the indicators of interest, construct statistical model with internet search results(ex. Google trends) and the existing survey statistics to predict current and future values of the indicators
	Sentiment analysis	• Analyze texts from SNS and determine polarity and emotional state of the writer, quantify the result to develop confidence/sentiment indicators
Seasonal adjustment		• Remove seasonal fluctuation factor to identify trends
Index calculation		• Calculate the indices with collected data

Vision	“Scientific administration based on data technology”			
Strategy	Implement big data governance	Publish statistics using big data	Construct statistical production support system	Reinforce internal competence
Tasks	1 Provide institutional framework to promote big data use	3 Survey and verify statistical demand on big data use	7 Construct and operate statistical production support system	9 Reinforce internal competence by big-data education
	2 Construct interagency cooperation system	4 Develop statistical index using big data	8 Provide and promote statistics produced using big data	
		5 Develop statistical production methodology using big data		
		6 Quality control and privacy protection for big data use		

STATISTICS KOREA

Thank you.