Estimating Dynamic Workplace Capacities by Means of Public Transport Smart Card Data and Household Travel Survey in Singapore

Sergio Arturo Ordóñez Medina and Alex Erath

The number and the temporal and spatial distribution of work locations are crucial information for any transport demand model. To generate the initial transport demand of MATSim, an activity-based multiagent simulation framework, it is necessary to determine dynamic workplace capacities with high spatial resolution, either on a parcel or even a building level. Commonly applied methods to derive work locations are based on census of enterprises information, unemployment insurance database, or combined information of a building’s gross floor area and individual work space requirements. As an alternative, the authors present a methodology that combines public transport smart card transaction data, travel diary surveys, and building information data sources. Work activities are detected from smart card transactions based on observed activity duration and start time and therefore related to public transport stops. To link the observed work activities to individual buildings, a linear programming optimization technique is applied that minimizes the walking time between public transport stops and potential work locations. The method classifies work activities in representative work schedules obtained by clustering methods. Information on maximum allowed building gross floor area derived from land use regulation is combined with estimates on individual work space requirements to ensure that buildings are only assigned with work activities according to their maximal capacity. To account for private transport based work activities, mode shares as observed in a travel diary are taken into account. To demonstrate the applicability, the proposed approach is implemented in Singapore and the results critically reviewed.

Commonly, authorities derive the spatial distribution of work locations either from an enterprise census or from estimating the number of employed persons based on information on the built environment. The first comes usually with the shortcoming that work location can be different from the employer’s office location which is listed in the enterprise census. In the latter case, the number of employees per work location is derived by dividing the gross floor area of individual buildings by land-use specific work space requirement \( I \). Such estimates are subject to substantial variation and therefore might not be reliable, especially if considered on the level of individual buildings, because many factors influence the space that is required for an individual worker: type of activity within use type, working practices, size of premises, location, region, economic cycle, building age, energy efficiency, reliance on technology, length of occupation and type of tenure. A third approach which is primarily employed by transport planners is to combine census data and transport surveys. In studies by McDonald and Prather \((2)\), McMillen \((3)\), García-López and Muñiz \((4)\), and others, employment density and decentralization is generated by combining fine grained information on individual work locations from travel surveys with population census data. However, the resulting spatial resolution is normally of aggregate nature, with no more than 100 zones for a large metropolitan area. Furthermore, all these methods fail to account for potential differences of work schedule characteristics among different work locations.

At the same time, new data sources that track the movement of individuals are emerging. Data generated by smartphones or other devices with Global Positioning System sensors, card transactions, or road pricing controls are examples of passively collected information that allow researchers to derive individual travel patterns and can serve as a source to impute work locations. In particular, smart card data have been analyzed in transport related studies now for more than 10 years. Chakirov and Erath \((5)\), to discover public transport users’ behaviors of Singapore; Seaborn et al. \((6)\), to analyze the multimodal travels in London; and Munizaga and Palma \((7)\), to estimate disaggregated origin–destination matrices in Santiago, Chile, demonstrate the potential of this comprehensive source of information. More recently, Devillaine et al. \((8)\) and Chakirov and Erath \((9)\) presented techniques for recognizing activities, in particular work activities, by reconstructing activity chains of public transport users. While such data sources can contain information for up to 97% of all public transport trips, it usually lacks information on the sociodemography of the traveler and, more important, covers only one mode of transport \((5)\).

Besides many other applications, the spatial distribution of work locations is a crucial data input for any transport demand model. For generating the transport demand of MATSim, an activity-based multiagent transport simulation framework \((10, 11)\), it is necessary to determine work locations with high spatial resolution, either on a parcel or even a building level. Besides the number of persons employed at different work locations, the MATSim framework allows incorporation of additional information on the characteristics of the work location such as typical occupation types and prevalent working schedules.

This paper presents a method that makes use of public transport smart card transaction data by combining it with a conventional travel

---

Future Cities Laboratory, Singapore ETH Centre for Global Environmental Sustainability, 05-01 CREATE Tower, 1 CREATE Way, Singapore 138602. Corresponding author: A. Erath, erath@v.t.baug.ethz.ch.

Transportation Research Record: Journal of the Transportation Research Board, No. 2344, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 20–30. DOI: 10.3141/2344-03
diary survey and the official land-use plan to impute the capacity of work locations on a building level. In addition, by including the observed time dynamics of public transport usage, information on prevalent work schedules is derived and attributed to individual work locations. Data from a travel diary survey is then used to account for spatial variations of modal share and work type characteristics. The objective of this innovative method is to obtain a better estimate of workplace capacities together with work schedule characteristics. It is organized as follows. After a description of the input data sets, the method is presented in detail and described in three steps. First is a description of how the source data (i.e., travel diary surveys, land use plans, and smart card transaction data) are combined. Second, observed public transport work activities are assigned to individual parcels on the basis of a closed-form optimization approach. Third, an inflation attributable to the mode share is performed, and employment locations are divided within parcels to single buildings. The application of the method for the case of Singapore is followed by conclusions from this case study and potential future improvements are discussed.

INPUT DATA

Public Transport Smart Card Data

In many cities, public transport smart cards have been introduced to replace or complement other forms of fare collection. As an additional benefit, detailed spatiotemporal data on usage patterns of public transport services is acquired, in particular if transaction data are collected for both boarding and alighting. Following their inception, smart card systems quickly became the dominant mode of payment for public transport services in Singapore, for example, for about 97% of all trips covered. Consequently, transaction records of such systems are extremely rich data sets for transportation research. The range of potential applications for research is vast and stretches from reconstructions of vehicles trajectories to the observation of the number of usage levels or crowdedness measurements of the systems.

The main disadvantage is the lack of sociodemographic information on the individual user because of privacy issues. However, in some systems, certain user groups use different card types for fare reduction purposes.

Travel Diary Survey

Travel diary surveys are among the most commonly used data sets for transport planning purposes. In such surveys, detailed information on individual travel behavior is collected based on a diary of journeys and activities performed on a given day. Because of intensive costs and labor, the sample size of travel diary surveys is usually restricted. Collected in small exercises annually or larger efforts every 4 to 6 years, such surveys usually cover several thousand records for a dedicated region and result in coverage of about 1% of the total population.

Geographic Information on Public Transport Stops, Land Use, and Building Footprints

Geolocation of public transport stops and potential work facilities, generated by combining building footprints and land-use information, is required to determine spatial relations between these facilities. Master plans usually indicate, on different scales, allowed land use and also allowed floor area ratios. The smallest area of consistent land use registration must be used to be related in the same scale with public transport stops as it is shown in Figure 1. Information on individual buildings derived from its footprints and height can be included to disaggregate even more of the estimation.

METHODOLOGY

The procedure was designed to estimate workplace capacities at a building level for a whole city by using public transport smart card data and a household travel survey. The method also classifies workers by their work start time and duration. Figure 2 presents how the input data are organized to feed the optimization model and how this result is processed to obtain the desired final distribution.

Preprocessing

Detection of Work Activities

As the sequence of trips that each public transport user performs in one day can be obtained with the smart card data, the time between two consecutive trips can be understood as an activity.

To impute the activities of a particular person from such data, the recorded daily journey chain of this person needs to be consistent. Consistency, in the context of smart card data records, means that the person who arrived at the activity location by public transport has to leave it after ending the activity also by public transport, otherwise the activity duration cannot be extracted. The most obvious cases of inconsistency can be identified by analyzing the distances between the alighting location of last journey and the boarding location of the following journey. In inconsistent consecutive public transport journeys, it is obvious that the person used another means of transport (e.g., car or walking). Although it is assumed that only one activity was performed between two consistent consecutive public transport journeys, it is not necessarily guaranteed. The same definition for activity consistency needs to be employed both for the detection model using travel diary survey data and application to public transport smart card data.

Clustering of Reported Work Activities and Classification of Detected Work Activities

Two-dimensional points of start time versus duration of reported work activities from the travel diary survey are taken as the input data of the clustering process. Specifying a desired number of clusters, a heuristic algorithm for the K-means problem is used, minimizing the distances to the cluster centroids through iterations. A principal-component analysis is applied to the input points, transforming them to a less correlated two-dimensional space, following which the resulting centroids are transformed back to the original start time–duration space. Each point is weighted to represent the importance of the record. The centroids of the clusters are the start time and duration of representative work schedules.

After applying the work activity detection model to a smart card transaction, data information on the start time and duration of work activities are derived. Thus, each detected work activity is classified in a certain work schedule by using Euclidean distances in the start
time–duration space. In this way, the number of workers in each stop (s) and at each work schedule (w) is obtained as $q_{sw}$.

Spatial Distribution of Travel Modes to Work

To account for spatial differences in mode choice of work journeys, it is proposed to estimate mode shares on a zonal level with travel diary surveys. Thereby, one needs to trade off between keeping sufficient observations to ensure reliable inflation factors and keeping a high level of spatial granularity. Starting with a fine grained zonal system (e.g., as used in aggregated transport demand models), an iterative procedure can be employed to merge zones gradually so that each new zone would feature a certain minimum number of observations. For each zone a mode share is obtained according to Equation 1. This same rate is assigned to the land use parcels that are contained in the zone (Equation 2).

$$n_{PuCons} + n_{Incons}$$

$ms_z = \frac{n_{PuCons}}{n_{PuCons} + n_{Incons}}$  \hspace{1cm} (1)

$$ms_p = ms_z \text{ if } p \subset z$$  \hspace{1cm} (2)

where

$ms_z = \text{mode share of zone}_z$,

$ms_p = \text{mode share of land use parcel}_p$,

$n_{PuCons} = \text{number of work activities featuring consistent public transport trips}$, and

$n_{Incons} = \text{number of work activities featuring inconsistent or private transport trips}$.

Maximum Number of Workers

Urban land use maps commonly specify the maximum allowable built area at parcel level. According to the type of work that people do, it is possible to estimate the area that those workers need. For example, for business type, a worker needs the average size of an office, but for industrial type, a worker uses a bigger area. By dividing the allowed built area by the type of worker area, an estimated value of...
the maximum number of workers in each land use parcel is obtained. This value is deflated in relation to the mode share value of the correspondent land use parcel:

\[ m_p = \left( \frac{a_p m_{p\text{, ms}}} {l_t} \right) \quad \text{if } p \text{ is of type } t \]  \hspace{1cm} (3)

where

- \( m_p \) = maximum number of workers in land use parcel \( p \) from public transport,
- \( a_p \) = maximum allowed area of land use parcel \( p \), and
- \( l_t \) = work space of type \( t \) workers.

**Travel Times Between Stops and Land Use Parcels**

Given a certain land use parcel, it is likely that a public transport worker arrives from nearby public transport stops. This relationship is measured through the walking travel time. For simplicity a maximum walking time is fixed, defining also that from a certain stop, if a land use parcel cannot be reached within this time the pair is not related. It is possible that the sum of all the maximum capacities of the land use parcels related to a given stop, is less than the number of people that arrive to the stop according to the smart card data. It means that some users of the public transport use different modes to reach their workplaces (taxi, shuttle bus, etc.). For those stops, farther land use parcels are related, supposing that those can be reached by motorized modes. In conclusion, this procedure measures the public transport accessibility of the land use parcels. More accurate results are achieved obtaining shortest paths with a high-resolution network to calculate walking or motorized travel times: \( t_{tp} \) (travel time from stop \( s \) to land use parcel \( p \)).

**Optimization Model**

Because of the various interconnections of land use parcels and public transport stops, the problem of assigning a distinct number of workers from each stop to its related parcels requires solving for the whole city in the same process. Although the number of workers would preferably be treated as integer values when determining the number of workers per land use parcel, instead of mixed-integer programming, linear programming techniques were used to restrain the complexity of the problem. However, resulting rational numbers indicating the number of workplaces for each parcel are not a real disadvantage. First, the subsequent inflation procedure relies on rational mode share indicators. Second, rational numbers work equally well as a measurement of attraction when used for trip assignment purposes.

The basic principle for assigning detected work activities from stops to land use parcels is to minimize travel times between those locations constrained to parcel capacities and potential interconnections.
between stops and parcels. The objective function given by Equation 4 is modeled along those lines.

$$\min \sum_{i \in I} \sum_{j \in J} \sum_{w \in W} x_{ijw} u_{wj} \quad 0 \leq x_{ijw}$$  \hspace{1cm} (4)$$

where

$x_{ijw}$ = number of workers from transit stop $i$, in parcel $j$, at schedule $w$;

$tt_{ijw}$ = travel time from stop $i$ to land use parcel $j$;

$S$ = set of public transport stops;

$P$ = set of land use parcels; and

$W$ = set of work schedules.

The first constraint, formulated in Equation 5, ensures that the total number of work activities $q_{sw}$ detected at a given stop $i$ with schedule $w$ is consistent with the number of assigned workers originating from stop $i$ with schedule $w$ when summed up over all land use parcels $j$:

$$\sum_{j \in J} x_{ijw} = q_{sw} \quad i \in S; \quad w \in W$$  \hspace{1cm} (5)$$

The second restriction sets for each land use parcel $j$ a maximum of the total assigned work activities $m_j$:

$$\sum_{i \in I} \sum_{w \in W} x_{ijw} \leq m_j \quad p \in P$$  \hspace{1cm} (6)$$

Postprocessing

Mode Inflation

The objective of the mode inflation process is to account for the unobserved share of workers in each land use parcel (e.g., those who do not consistently use public transport). As described earlier, each land use parcel is exclusively related to a zone $z$ of constant mode share $m_z$. After all public transport work activities $x_{wz}$ within all land use parcels related to a single zone are summed (Equation 7), the number of unobserved work activities of a particular work schedule $y_{wz}$ is calculated by using Equation 8:

$$x_{wz} = \sum_{i \in I} x_{ijw}$$  \hspace{1cm} (7)$$

$$y_{wz} = \sum_{j \in J} x_{wz} \left( 1 - \frac{m_z}{m_j} \right)$$  \hspace{1cm} (8)$$

Thus, it is assumed that the relative proportion of different schedule types within each zone of constant mode share as observed from public transport smart card records apply also for all inflated work activities.

Distribution of the inflated work activities in relation to the remaining capacities of all parcels within a zone ensures that work activities not consistently observed by public transport smart card records are referred to parcels with lower public transport accessibility. This condition is expressed in Equation 10, which defines the proportion of unobserved work activities for each land use parcel $r_p$ within a zone:

$$r_p = \frac{m_z - x_{wz}}{\sum_{j \in J} (m_z - x_{wz})} \quad \text{if } p \subset z$$  \hspace{1cm} (10)$$

The number of work activities $y$ of schedule type $w$ within each land use parcel $p$ can then be obtained with Equation 11.

$$y_{pw} = y_{wz} r_p$$  \hspace{1cm} (11)$$

Distribution to Single Buildings

Within each land use parcel, work activities are attributed to single buildings in relation to the imputed gross floor area and the number of stories of each building:

$$n_{sw} = \frac{a_b f_b}{\sum_{j \in J} (a_j f_j)} \left( x_{wz} + y_{wz} \right) \quad \text{if } b \subset p$$  \hspace{1cm} (12)$$

where

$n_{sw}$ = total number of workers in building $b$ at schedule $w$,

$a_b$ = footprint surface of building $b$, and

$f_b$ = number of stories of building $b$.

IMPUTING WORK LOCATION ON LEVEL OF INDIVIDUAL BUILDINGS

Input

Travel Diary

Singapore’s Land Transport Authority conducts a travel diary survey, Household Interview Transport Survey (HITS), every 4 years. In 2008, HITS covered approximately 1% of the population—around 38,000 participants within 10,500 households, including citizens, permanent residents, and legal immigrants—which represented comprehensively the 4.839 million legal residents of Singapore in that year. The survey is conducted on a household level and contains, for all household members, their sociodemographic characteristics and the information of all activities that required a trip of more than 500 m performed over a day.

Public Transport Smart Card Transactions

In 2002, a smart card–based automated fare collection system was introduced in Singapore for all public transport services. The system records every trip of every public transport user in the city. Adults and senior users record the beginning and the end of the trip by tap-in and tap-out activities. Children and high school and undergraduate students can obtain concession passes for which only tap-out is recommended but not enforced.

Land Use Information

Land use information is obtained from the 2008 master plan published by Singapore’s Urban Redevelopment Authority (14). For 10,762 individual land-use parcels that are categorized by
31 land use types, the master plan contains information on the type of allowed land use and gross plot ratio (gross floor area of buildings within a parcel divided by the total surface of a parcel). Footprints of 159,374 buildings for 2008 are provided by the Singapore Land Authority.

Preparing Input Data

Detection of Work Activities from Public Transport Smart Card Transactions

Chakirov and Erath (9) presented a discrete-choice model designed to detect work activities. The model combined activity start time, duration, and land-use data and was estimated by using HITS 2008. When the model parameters were applied to a holdout sample, the model successfully recognized 97.5% of all cases.

Chakirov and Erath (9) applied the parameter estimates of the model to public transport smart card data collected in 2011. By running the model for each workday individually and averaging the results, potential day-to-day variations were controlled. With this procedure, 717,740 of 2,366,686 observed activities that fulfilled the consistency criteria were detected as work activities.

Clustering Work Activities in Relation to Start Time and Duration

Figure 3 shows the result of application of the algorithm discussed in the earlier section on preprocessing to the work activities reported in HITS 2008, aiming at 10 clusters. The size of each point represents the trip factor, and the less transparent clusters include more records. The shape of the clusters shows that the clustering process was performed in a rotated two-dimensional space. Therefore, work activities detected with the previous procedure were assigned to one of those clusters in relation to observed start times and activity duration.

Spatial Distribution of Travel Modes to Work in Singapore

To account for spatial differences in mode choice of work journeys, mode shares were calculated on a zone level as presented in Chakirov and Erath (9). As HITS counts only about 13,000 work activities, many of which are accumulated in the central business district, the HITS sample does not provide sufficient observations for a reliable mode share calculation in each of the 1,092 traffic planning zones used in the transport demand model of the Land Transport Authority. Therefore, adjacent traffic planning zones were merged by using an iterative procedure and a threshold of at least 50 HITS work activity observations per zone. As a result, 162 aggregated zones were obtained. Some few exceptions containing a smaller number of observations were zones without any adjacent zones, as in the case of some islands.

To define a mode of each work activity, journeys to and from work were considered jointly and only one mode was assigned to each work activity. Thereby, only two modes were differentiated. First, the public transport mode included all work travels, which form a consistent public transport journey chain encompassing a work activity that can be observed from public transport journey records. Second, all other work travels consisting of journey chains that would be identified as inconsistent or would not be visible at all within the public transport smart card record were considered private or other transport mode. This system of two modes should allow the inflation of the majority of work activities that cannot be detected.
Maximum Number of Work Activities per Land Use Parcel

To calculate the maximum capacities of work activities in each land use parcel, the following assumptions were considered:

- Gross floor area in buildings within each land use parcel was given by the parcel’s area multiplied by the maximum allowable gross plot ratio.
- In accordance with allowed types of development within each land use type, assumptions on the general work space requirements were derived from the UK Homes and Communities Agency (1). Table 1 presents the employed work space requirements.
- For each parcel, the maximum capacity of work activities was partitioned in relation to the mode share of the related (aggregated) traffic planning zone.
- A white-type parcel in Singapore means that the planning intention of the parcel is open to many uses, including residential, office, and recreational. White-type parcels were analyzed individually and categorized with another type according to the location and the neighborhood.

Travel Times Between Public Transport Stops to Land Use Parcels

With 4,397 public transport stops and 10,762 parcels, 47,320,514 potential relationships needed to be considered when work activities from stops to parcels were being assigned. Travel times between stops and parcels were computed on the basis of the shortest-path routing by using a high-resolution road network featuring 43,118 nodes and 79,835 links that was provided by Navteq. Because this network did not cover all pedestrian infrastructure, such as underpasses, stops were related to the three nearest links and the shortest-path calculation was performed to all links within a land use parcel. Of all the resulting potential connections, the one with the least travel time was selected. However, if based on the euclidean distance and under the assumption of a walking speed of 4 km/h, the maximum walking time exceeds 15 min, the routing is omitted, and the travel time of the stop-to-parcel relationship is set to 600 min. When the number of work activities detected from analysis of public transport smart cards exceeds the maximum capacity of the related parcels, the process is repeated by assuming a travel speed of 16 km/h [as the public transport mode speed specified in Lee et al. (15)] to cover motorized travel by taxi or shuttle bus. On average, 40 work areas were related to each stop, and 16 stops were related to each work area. The computation time to perform this task on an up-to-date personal computer (Intel i7 processor with four cores clocked at 2.7 GHz and 8 GB of RAM memory) is about 30 min. The results of this process are presented in Figure 4, which shows all considered stop-to-parcel relationships.

Obtaining Workplace Capacities

Solving Optimization Model

By combining 47,320,514 potential relationships and 10 work schedule clusters, the number of decision variables in the optimization model amounts to 473.2 million. The set of constraints is given by the number of stops multiplied by the number of considered work schedules, which amounts to 43,970, and the maximum capacities is 10,762 land use parcels. To solve the linear program, the CPLEX software package was used on a server with 72 cores and 512 GB of RAM memory, solving the optimization model in approximately 3 h by using more than 10 cores and more than 200 GB of RAM. The result of the procedure is the allocation of work activities observed from the data public transport smart cards at the level of bus stops to land use parcels, as shown in Figure 5.

Mode Inflation

Mode inflation is performed on the level of 162 aggregated traffic zones as used in Chakirov and Erath (9). The total number of added work activities within each zone is then broken down to individual

| Type                        | \(m^2/\text{Worker}\) | Type                        | \(m^2/\text{Worker}\) |
|-----------------------------|-----------------------|-----------------------------|-----------------------|
| Residential                 | 10^4                  | Place of worship            | 100                   |
| Residential, 1st floor commercial | 250                  | Civic and community institution | 16                   |
| Commercial and residential  | 250                   | Open space                  | 10^5                  |
| Commercial                  | 18                    | Park                        | 10^5                  |
| Hotel                       | 500                   | Port or airport             | 36                    |
| Utility                     | 60                    | Beach area                  | 10^5                  |
| Business park               | 10                    | Sports and recreation       | 100                   |
| Business park, white        | 12                    | Water body                  | 10^5                  |
| Business 1                  | 13                    | Road                        | 10^5                  |
| Business 2                  | 36                    | Transport facilities        | 70                    |
| Business 1, white           | 16                    | Rapid transit               | 36                    |
| Business 2, white           | 40                    | Cemetery                    | 10^4                  |
| Residential, institution    | 16                    | Agriculture                 | 10^4                  |
| Health and medical care institution | 16                  | Reserve site                | 10^4                  |
| Educational institution     | 25                    | Special use                 | 10^4                  |
land use parcels. When doing so, the amount of yet unallocated capacities within each land use parcels is considered. For the newly allocated work activities no further information on underlying work schedules is available. Therefore, work schedules of newly allocated work activates within each parcel are attributed proportionally to the work schedules assigned with the minimization model.

Distribution to Individual Buildings

Within each land use parcel, all work activities and corresponding schedules are distributed to 158,204 single buildings. Since besides the footprint no further information was available, only the surface of the building footprints could be considered as weight for each building. Or expressed in other words, it is assumed that each building within a given parcel has the same number of stories. Figure 6 presents total capacities of all kinds of workers on a building level in the whole island, and the distribution of the defined schedules on individual buildings is detailed for the southern part of Singapore.

Assessment of Results

Lacking data sources on the actual distribution of work locations in Singapore, the model can currently only be validated by assessing different result visualizations. As shown in Figure 6, generally the presented methodology is able to generate a plausible distribution of work locations: main centers of employment such as the central business district at the southern tip of the island and the industrial estates in the west are characterized well both in terms of the number of work locations and density. The variation on the level of individual buildings seems plausible as well, in particular in the city center, the distribution of work locations reflects well the variety of building typologies. However, an analysis of outliers reveals also potential shortcomings of the method and the data employed, but also shows that better results can be achieved with reasonable additional efforts.

The public transport smart card data were collected in April 2011, while information on the built environment dates back to the year 2008. Due to the rapid development of Singapore during those years this caused some inconsistency. An example of such a mismatch is Marina Bay Sands complex, an integrated resort featuring a hotel of 2,561 rooms, a shopping center with more than 300 stores and leisure attractions such as a theatre and casino opened in 2010. Since the Marina Bay Sands complex is not covered in the land use information from 2008, observed work activities from nearby public transport stops have not been assigned to other buildings, a result which is reflected in substantially longer walking distances.

The assumption to minimize only the walking time between the alighting stop and the final destination might not hold thoroughly valid as people plan their routes by aiming to minimize the generalized cost of the whole trip. When the methodology is applied to Singapore, such a mismatch is observable for the case of a centrally located mass rapid transit (MRT) station, City Hall. Despite its central location, the buildings within the immediate surroundings are of comparably small size, hosting potentially a rather low number of work locations. However, the MRT station is linked through an underground mall of almost 1 km length with numerous shops to two of the largest commercial developments in Singapore named Suntec City (shopping mall, office towers, and convention center).
FIGURE 5  Work activities as observed from public transport smart card data assigned to land use parcels for Singapore.
and Marina Square (shopping mall with 3 major hotels attached). Protected from outside environment, commuters apparently prefer to pass through this link rather than transfer to a bus or MRT that would bring them closer to their destination.

The case of the Singapore Zoo is an example of an outlier detected through the analysis resulting distances between work location and public transport stop. According to its land use designation and gross surface area, the maximal number of work activities is set too low because the assumed work space requirement for a park does not reflect well the special case of a zoo. Surrounded by other parcels featuring low work density land use types, detected work activities need to be distributed to parcels which are unrealistically far away. Obviously, in this particular case, the problem can be eliminated by changing the work space requirement for the land-use parcel of the zoo and rerunning the process. In other cases, similar situations might actually reflect reality, for example, commuters using a company shuttle bus between an MRT station and a work facility.

Therefore, revisiting such outliers and assessing the actual location situation are advisable.

**CONCLUSION AND FUTURE WORK**

An innovative method for estimating workplace capacities of a city was designed, implemented, and tested within the city of Singapore. The method combines information from several data sets, such as a household travel survey and public transport smart card transaction data. The method categorized workers by schedule allowing dynamic demand generation and distribution. As the aim of the project was to feed an agent-based and activity-based transport model (MATSim-T), the level of spatial granularity is an individual building. The core of the process is an optimization problem to minimize the travel time of workers after leaving public transport stops. Due to the size of the problem, 47.32 million decision variables must
be solved, thus requiring a computer with more than 30 cores and more than 200 GB for solving in about 3 h.

In absence of adequate data for validation, future work should address the sensitivity of the proposed methodology to its key assumptions: first, the methodology should be tested to determine what extent the results are dependent on the work space requirements for individual land use types. Second, it was assumed that all land use parcels are fully developed. Since such information is only used to set upper limits of the number of work activities, but the distribution of work activities is more related to the public transport smart card observations, one might argue that the actual implications are only of secondary order. However, in the case of Singapore, data on building sizes are publicly available for a restricted downtown area through Google Earth. By rerunning the process including such data and comparing it with the results presented in this study, one could test the sensitivity of the method towards the influence of such information. Furthermore, the model should be validated on a more aggregate level by comparing the results if workplace information on the level of administrative zones becomes available from other sources such as the tax revenue authority or business census.

The mode inflation process implicitly assumes that mode share ratios for a zone over 24 h remains constant. Considering potential differences for off- and on-peak, as well as for night-shift workers, such an assumption might be simplistic and affect the result in terms of the effective distribution of prevalent schedules within each zone, while the total number of workplaces remains unaffected. A further refinement in that direction would require either more extensive travel diary survey coverage or application of advanced statistical inference techniques such as Gibbs sampling.

It is argued that shortcomings presented in the assessments of results can be approached with reasonable additional efforts to improve the results: using different data sets which are consistent with respect to the survey year could prevent problems as observed in the case of Marina Bay Sands. By including also the information on the boarding station, the optimization model could be reformulated by minimizing generalized cost. Although this reformulation would come at the cost of a substantial increase of the computational intensity, it should be feasible. In addition, employing information on the pedestrian network would improve the result in that direction as well, but such information is still hard to obtain. Finally, special land use situations such as zoos, airports, or exhibition centers can be manually studied in order to assign reasonable maximum work capacities.

If a mixed-integer programming solution is desirable to estimate integer capacities, or if computation power is not available, a spatial clustering of the stops can be performed for solving many small problems with the same strategy. However, for work areas related to stops in more than one cluster (areas in the intersection of the clusters), a distribution of their maximum capacities would need to be performed beforehand according to a further set of assumptions, for example, travel times to the stops that belong to each cluster (assuming that clusters with more accessible stops to an area deserve more of its capacity), the number of workers that arrive to the stops (assuming that clusters with more crowded stops deserve more of capacity), or a combination of both.

ACKNOWLEDGMENTS

The authors thank the following Singapore authorities for providing access to data and valuable review: Land Transport Authority, Urban Redevelopment Authority, and Singapore Land Authority (SLA Digitised Land Information). Financial support came from the National Research Foundation of Singapore and ETH Zurich research fund.

REFERENCES

1. Employment Densities Guide, 2nd ed. UK Homes and Communities Agency, London, 2010. http://www.homesandcommunities.co.uk/download-doc/6155/10397.

2. McDonald, J. F., and P. J. Prather. Suburban Employment Centres: The Case of Chicago. Urban Studies, Vol. 31, No. 2, 1994, pp. 201–218.

3. McMillen, D. P. Employment Densities, Spatial Autocorrelation, and Subcenters in Large Metropolitan Areas. Journal of Regional Science, Vol. 44, No. 2, 2004, pp. 225–243.

4. Garcia-López, M. A., and I. Muñiz. Employment Decentralisation: Polycentricity or Scattering? The Case of Barcelona. Urban Studies, Vol. 47, No. 14, 2010, pp. 3035–3056.

5. Chakirov, A., and A. Erath. Use of Public Transport Smart Card Fare Payment Data for Travel Behaviour Analysis in Singapore. Presented at 16th International Conference of Hong Kong Society for Transportation Studies, Hong Kong, China, Dec. 2011.

6. Seaborn, C., J. Attanucci, and N. H. M. Wilson. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. In Transportation Research Record: Journal of the Transportation Research Board, No. 2121, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 55–62.

7. Munizaga, M. A., and C. Palma. Estimation of Disaggregate Multimodal Public Transport OD Matrix from Passive Smart Card Data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, Vol. 24, 2012, pp. 9–18.

8. Devillaine, F., M. A. Munizaga, and M. Trépianer. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. In Transportation Research Record: Journal of the Transportation Research Board, No. 2276, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 48–55.

9. Chakirov, A., and A. Erath. Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport. Presented at 13th International Association for Travel Behaviour Research, Toronto, Ontario, Canada, July 2012.

10. MATSim, Multi-Agent Transportation Simulation. http://www.matsim.org.

11. Balmer, M. Travel Demand Modeling for Multi-Agent Traffic Simulations: Algorithms and Systems. Ph.D. dissertation. Swiss Federal Institute of Technology in Zurich, Switzerland, May 2007.

12. Sun, L., D.-H. Lee, A. Erath, and X. Huang. Using Smart Card Data to Extract Passenger’s Spatio-Temporal Density and Train’s Trajectory of MRT System. Presented at the ACM SIGKDD International Workshop on Urban Computing, Beijing, Aug. 2012.

13. Eom, J. K., M. H. Choi, and L. Lee. Evaluation of Metro Service Quality Using Transit Smart Card Data. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.

14. The Planning Act: Master Plan Written Statement. Technical report. Urban Redevelopment Authority, Singapore, 2008.

15. Lee, D.-H., L. Sun, and A. Erath. Study of Bus Service Reliability in Singapore Using Fare Card Data. Presented at 12th Asia Pacific ITS Forum & Exhibition, Kuala Lumpur, Malaysia, April 2012.

The Transportation Demand Forecasting Committee peer-reviewed this paper.