Physica A 557 (2020) 124971

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Thermodynamics of urban growth revealed by city scaling

Lorraine Sugar [a], Christopher Kennedy [b,*]

[a] Department of Civil Engineering, University of Toronto, Canada
[b] Department of Civil Engineering, University of Victoria, Canada

ARTICLE INFO

Article history:
Received 9 March 2020
Available online 27 July 2020

Keywords:
Cities
Scaling
Thermodynamics
Maximum Power Principle

ABSTRACT

Cities are complex systems, where micro-scale phenomena lead to emergent, macro-scale patterns. Scaling in cities – the study of how characteristics of cities change with urban population – has been explained based on social interactions and networks, but cities are also governed by the laws of physics, such as thermodynamics and conservation of mass. Here we explore scaling laws in the context of the city as a far-from-equilibrium thermodynamic system, showing how scaling phenomena are consistent with ideas of cities as dissipative systems, the emergence of higher-order structures, and the Maximum Power Principle. We theorize that as cities grow, they use an increasing amount of energy and resources to produce an incremental change in useful work done. Furthermore, they require additional energy and materials to build higher-order structures, such as transportation systems, to overcome increasing density and congestion. We support our theoretical approach with new empirically observed scaling relationships.

City scaling laws [1–3] capture the predictability of emergent phenomena observed in urban systems. The scaling laws are empirically derived relationships between city size (i.e., population) and other extensive indicators, such as number of crime, GDP, length of infrastructure, and household electricity consumption, for a system of cities in a given country (i.e., a cross-section of cities). There is typically superlinear scaling of social quantities, including both desirable social quantities (e.g., wages, GDP, new inventions) and undesirable social quantities (e.g., crime, disease transmission). Physical quantities display sublinear scaling, including urban infrastructure, such as electricity network length and length of road surface. Quantities required for individual maintenance tend show linear scaling, such as total housing, total employment, and household electricity consumption. Together, these results indicate that in each system of cities certain individual needs remain consistent, while social effects exhibit agglomeration economies and physical infrastructure achieves economies of scale. By virtue of their interdependent systems, larger cities accomplish more socioeconomically with less physical infrastructure.

Bettencourt [3] developed a theoretical model describing the underlying city processes that result in the observed scaling relationships. The model, called *The Origin of Scaling in Cities*, combines spatial equilibrium from economic geography with physical properties of social and infrastructural networks, making it possible to derive scaling exponents from general first principles. One way to visualize the model is as a collection of social interactions occurring on a network, shown in Fig. 1. Each individual travels through the network in three-dimensional space, represented by their "tube". A social interaction occurs when tubes cross. Variables related to social interactions scale superlinearly, while variables

* Correspondence to: University of Victoria, PO Box 1700 STN CSC Victoria BC V8W 2Y2, Canada.
E-mail address: cakenned@uvic.ca (C. Kennedy).

https://doi.org/10.1016/j.physa.2020.124971
0378-4371/

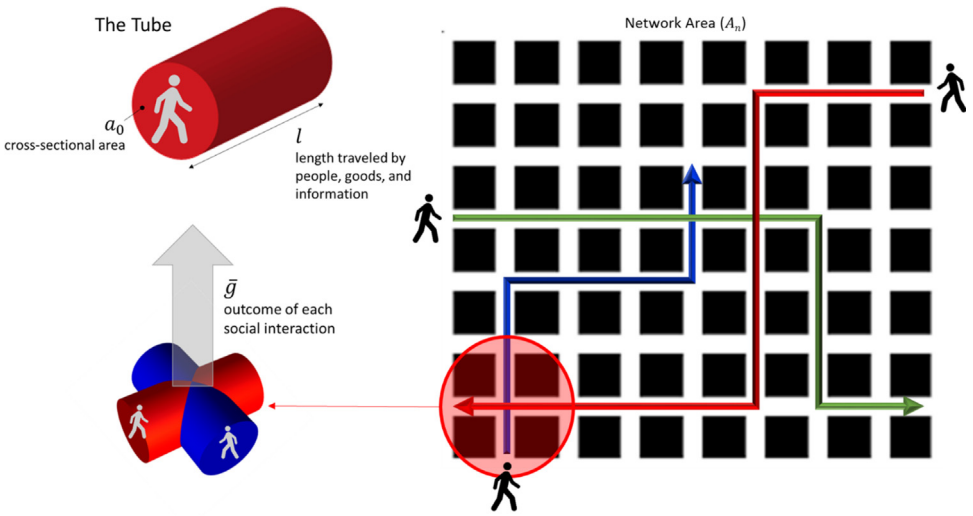

Fig. 1. Conceptualization of *The Origin of Scaling in Cities* model [3]. Individual 'tubes' as they travel around the city on a network. When tubes cross, a social interaction occurs.

related to the network scale sublinearly. Bettencourt's theoretically derived scaling relationships fall close to those that are observed empirically.

In terms of energy-related variables, total energy consumption and electricity transmission losses scale superlinearly, as shown by Bettencourt et al. [1] for German cities. The authors make an important observation: the majority of systems in nature evolved to minimize wasted energy, while, in contrast, cities appear to have evolved in such a way that maximizes energy consumption and loss. In this sense, the city is a *dissipative structure*. Coined by Prigogine, a dissipative structure is a far-from-equilibrium system that exchanges matter and energy with its environment, increasing its internal complexity to do so more effectively [4]. Schneider and Kay [5] and Kay [6] apply the concept of dissipative structures to ecosystems, and Bristow and Kennedy [7] to cities. As they mature, ecosystems develop more complex structures with greater diversity and more cycling, which systematically increases the system's ability to use incoming solar energy [5,6]. Bristow and Kennedy [7] describe cities as open thermodynamic systems that exchange energy and matter with their surrounding environments. As cities grow, they increase in complexity by building infrastructure, economies, and social institutions where incoming energy and matter is stored and metabolized, and in turn expel waste in the form of energy (e.g., heat) and matter.

1. Thermodynamics and urban scaling

Our objective is to examine scaling patterns in thermodynamically related measures of cities, thereby contributing to the understanding of the physical constraints on the growth of cities. During the conceptual development of the laws of thermodynamics, Clausius described that heat (i.e., energy) undergoes various transformations when interacting with a mass, as shown in Eq. (1) [8]:

$$dQ = dH + dW + dI \tag{1}$$

The quantity of heat imparted on a body (dQ) is divided among internal heat or thermal content of the body (dH), external work against forces outside the body (dW), and internal work against forces of attraction between constituent molecules, such as phase change and creation of ordered structures (dI). The body in this context is a thermodynamic system with constituent molecules interacting and exchanging energy with each other and the external environment. According to the *Maximum Power Principle* [9], the most successful systems adjust items on the right side of Eq. (1) (dH, dW, and dI) in order to maximize incoming energy (dQ).

If we consider Clausius' "body" to be a city, we can assign analogous quantities to each term in Eq. (1). The heat imparted on a body (dQ) is imported energy; the internal heat or thermal content of the body (dH) is the energy consumed for everyday city life; the external work against forces outside the body (dW) is the energy required to produce exports; and the internal work against forces of attraction between constituent molecules (dI) is the energy required to construct and maintain the physical fabric of the city, i.e., the capital assets of buildings and infrastructure.

We can examine how cities of different sizes apportion their energy balance quantities by examining scaling relationships for each term in Eq. (1). Due to the complexity of energy flows through cities, we investigated the terms on the right-hand side of Eq. (1) using measures of economic activity. Cities import energy in forms of differing quality,

e.g., electricity, fuels, and natural radiation. They also import raw materials and consumer products that contribute to the activities in cities. Moreover, as the imported energy becomes transformed in cities, deciphering energy used in construction activities and in the export of goods and services from other forms of energy is complicated—and not currently possible. Economics has a clear and established method for designating value of exports and activity in various sectors of the economy, thus providing a source of data that is more appropriate for our purposes than proxy energy estimates.

As outlined above, Bettencourt et al. [1] have shown that total energy consumption (dH) is superlinear for cities in Germany. We might therefore initially expect that all of terms on the right-hand side of Eq. (1) to be superlinear. Using data for U.S. cities (see Appendix), we observe a superlinear scaling relationship for GDP (Fig. 2A), confirming a result obtained by Bettencourt. In Fig. 2B, we also show a superlinear scaling relationship for construction sector GDP, an economic measure for energy required to construct and maintain physical infrastructure (dI). For exports from US cities, however, we find there is a linear, or slightly super-linear, scaling relationship (Fig. 2C); $\beta \sim 1.04$, but it is not significantly different to 1.0. These results can be interpreted as follows: as cities grow, they consume even more energy per person and invest even more in maintaining and growing their physical infrastructure, while their exports per person remain relatively constant.

The result for exports is particularly interesting. There is no theoretical treatment of exports in Bettencourt's model, but the empirical result parallels observations made by Schneider and Kay [5] for dissipation of gradients in Bénard cells. Bénard cells, a classic example of dissipative structures, are a pattern of convection cells that form in a plane horizontal layer of fluid that is heated from below, creating a temperature gradient in the fluid. It becomes more difficult to maintain the temperature gradient as the system of Bénard cells becomes more organized, requiring an increasing amount of energy to incrementally increase the temperature gradient. In other words, an increasing amount of energy goes into maintaining and growing the cells' structure, leaving an incremental amount to "flow through" the system. Our results suggest a parallel for cities in that a disproportionately increasing amount of energy is required by growing cities to produce an additional incremental unit of exports.

2. Cities as dissipative structures

Construction activities in cities are inherently part of their growth process and give rise to areal expansion, densification and the formation of higher-order structures, such as transportation infrastructure. In thermodynamic terms, the construction activities build dissipative structures in cities.

By looking at certain scaling relationships, we can see the characteristics of cities of different sizes, particularly in how they behave as dissipative structures. Several scaling relationships described above are relevant in this context. For example, length of infrastructure and land area show sublinear scaling, while energy consumption and GDP show superlinear scaling. The simple interpretation is that larger cities achieve more energy consumption and GDP per capita with less physical structure per capita—i.e., they are more effective systems. We now show that this is a result of the adaptation of the dissipative structure to maximize energy input.

A theoretical scaling relationship for population density can be derived starting with Bettencourt's theoretical scaling relationship for land area:

$$A = aN^{\alpha} \tag{2}$$

where A is land area, a is a constant, N is population, and α is 2/3. Rearranging and simplifying yields the following theoretical scaling relationship for population density, D:

$$D = \frac{N}{A} = dN^{1-\alpha} \tag{3}$$

where $d = 1/a$. The theoretical scaling exponent for population density is therefore 1/3. Population density behaves as a physical quantity that exhibits sublinear scaling: there are economies of scale in terms of how many people can occupy the 'tubes' shown in Fig. 1. It is important to note that even though the scaling is sublinear, population density is increasing as cities become larger. Empirical scaling relationships for population density for U.S. cities vary with the delineation of city boundaries, ranging from 0.51 ± 0.07 for Metropolitan Statistical Areas (Fig. 3A) based on commuter sheds to 0.12 ± 0.03 for Urbanized Areas (Fig. 3B) based on the built environment (see Appendix for details).

One of the consequences of increasing population density is traffic congestion. Congestion is a form of internal disorder resulting from inefficient use of infrastructure, which causes waste of both fuel and time. Energy dissipated when the infrastructure network is congested does not lead to useful economic work or reinvestment in the system. The waste produced by congestion can be quantified in congestion costs to an economy, as done so for cities in the U.S. by Texas A&M Transportation Institute and INRIX [10]. In Fig. 2D, we show that congestion costs scale superlinearly, indicating that as cities grow, they experience even more internal disorder leading to energy and economic waste.

Dissipative structures adapt to increasing internal disorder by creating higher-order internal structures. For example, the Bénard cells described above are the macroscopic, ordered structures (i.e., convective cells) created following microscopic, random movement of molecules (i.e., conduction). Cities, as dissipative structures, build higher-order infrastructure in response to increasing internal disorder and congestion. In this sense, the physical manifestations of the observed

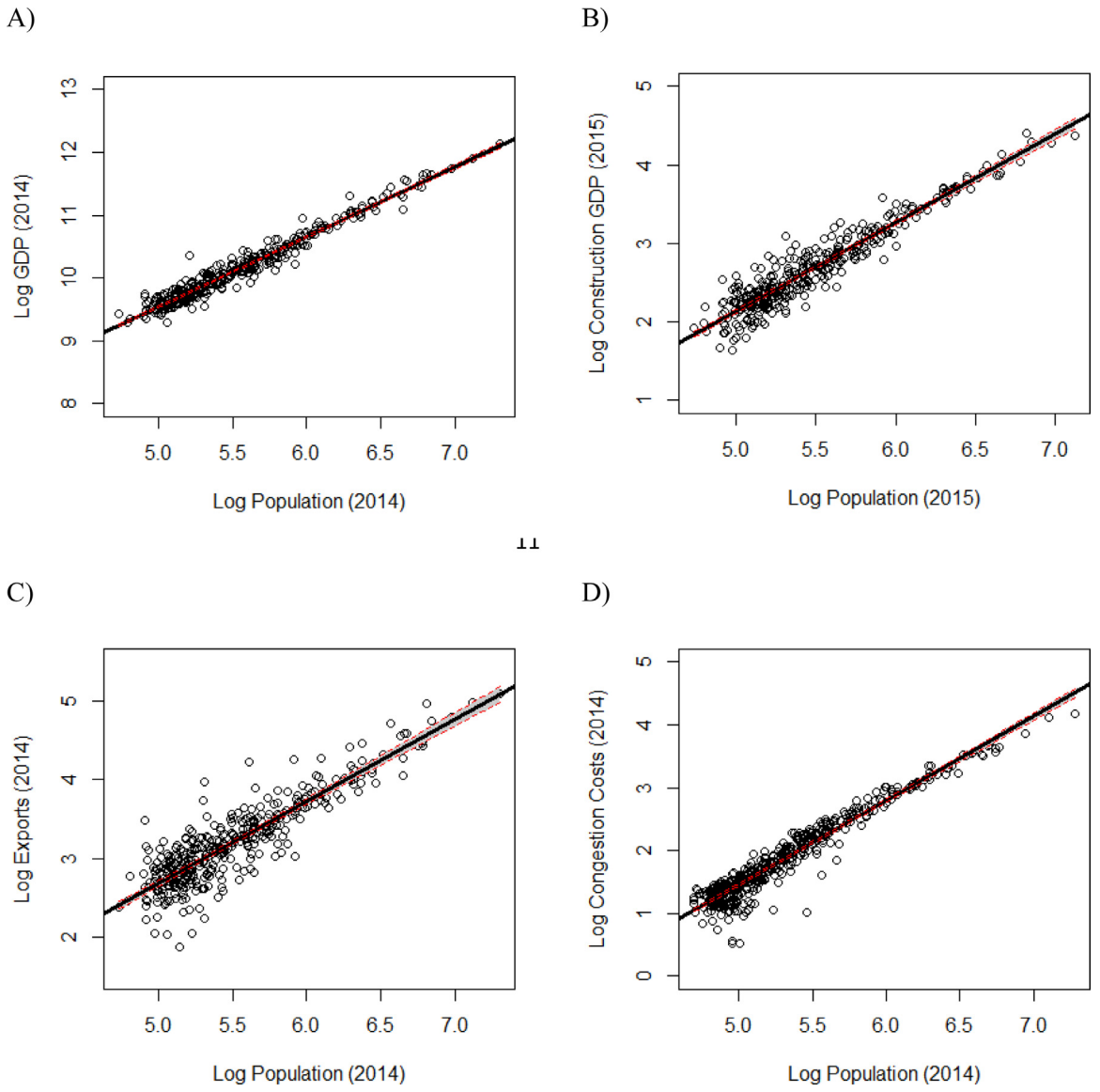

Fig. 2. Scaling relationships for: (a) GDP (real GDP in 2009 chained dollars) ($\beta = 1.11 \pm 0.02$, $R^2 = 0.96$, n = 364), data is for metropolitan statistical areas in the United States; (b) Construction GDP ($\beta = 1.12 \pm 0.04$, $R^2 = 0.90$, n = 341), data is for metropolitan statistical areas in the United States; (c) International exports ($\beta = 1.04 \pm 0.05$, $R^2 = 0.79$, n = 381), data is for metropolitan statistical areas in the United States; (d) Congestion costs ($\beta = 1.34 \pm 0.03$, $R^2 = 0.93$, n = 470), data is for urbanized areas in the United States.
Source: (a), (b) [11]; (c) [12]; (d) Texas A&M Transportation Institute and INRIX, 2015.

scaling relationship for population density are profound. Although the relationship is sublinear, population density increases as cities increase in population. Cities do not just expand outwards at uniform density but are built upwards with higher densities. In other words, we can say that larger cities build higher-order structures into their infrastructure – that is, increase the internal complexity of their transportation, social, and economic networks – in order to effectively accommodate more people. This is a feedback mechanism that enables increasing energy and economic inflows to the system, thereby fueling urban growth (Fig. 4).

We show the effects of higher-order internal structures empirically in Fig. 5 with data from cities in the U.S. We use passenger-kilometers-traveled (PKT) as an indication of internal activity, and the scaling relationship between population and PKT varies for cities with different types of transit systems. Cities with conventional lower-order bus systems show near-linear scaling; that is, the internal activity through the city grows more-or-less evenly with population. However, once a city has a higher-order of transit infrastructure, PKT scales superlinearly. There is more movement per capita in cities with light rail or bus rapid transit (with a scaling exponent of 1.28), and even more movement per capita in cities with hybrid rail or subway systems (scaling exponent of 1.66). Cities with higher-order internal structures have internal

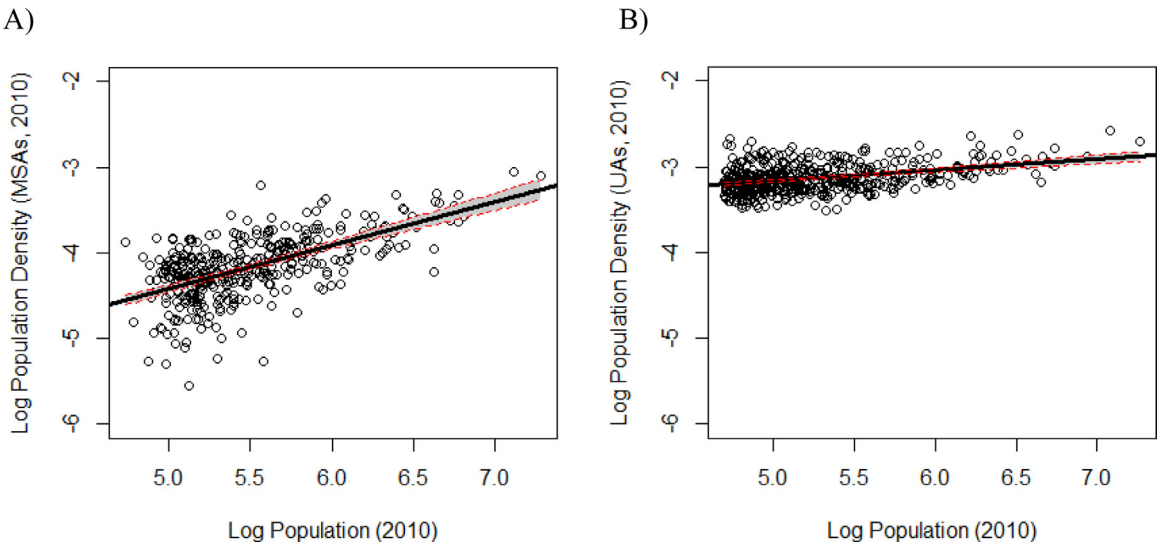

Fig. 3. Scaling relationships for population density: (A) data is for metropolitan statistical areas in the United States from the year 2010 ($\beta = 0.51 \pm 0.06$, $R^2 = 0.39$, n = 374); (B) data is for urbanized areas in the United States from the year 2010 ($\beta = 0.12 \pm 0.03$, $R^2 = 0.14$, n = 497). *Source:* [13].

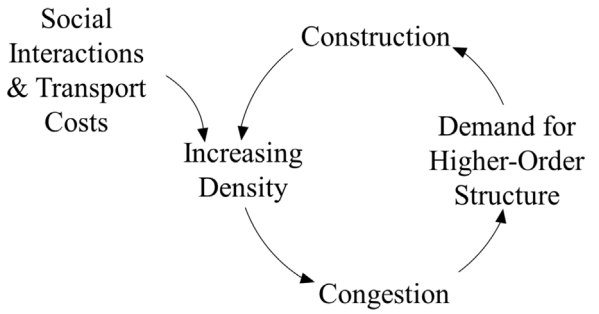

Fig. 4. Feedback processes leading to urban growth. Density is increased with increasing social interactions and transport costs. As density increases, there is more congestion and therefore demand for higher-order structures, such as transit connections. This leads to construction of infrastructure and buildings, which in turn allows for increasing density.

activity that scales even more rapidly than their counterparts. Internal activity is related to energy and economic inflow in terms of transportation energy used, but also more broadly because more internal movement in cities means more economic activity, i.e., more trips to businesses, more purchasing of goods and services, and more social activities.

3. Discussion

Our thermodynamic model for urban growth and new empirically observed scaling relationships are important for understanding the physical bounds to urban sustainability. The observation that urban density increases with city-size, albeit sub-linearly as predicted by scaling theory, is important. It implies that densification will tend to occur with city growth, even in the absence of explicit policy. Moreover, the development of transit systems to relieve the naturally occurring congestion can be expected to occur in a similar way that dissipative systems create higher order structure.

The implications of the above findings are two-fold. First, the Maximum Power Principle – i.e., the most successful systems adjust their internal consumption and physical in order to maximize incoming energy [9] – indicates that efficiency measures in cities in fact drive energy inflow and economic growth. This phenomenon has been observed in other disciplines, notably the Energy Efficiency Rebound Effect, where households tend to use money saved on energy bills for other energy-intensive activities, such as purchasing more electronics or airplane travel. In contrast to natural systems, which maximize a limited supply of energy from the sun, human systems have adapted to reinvest stored energy into increasing the flow of energy into the system and developing new ways to consume it.

Second, from the scaling relationships described above, we can see that larger cities are most effective at the process of adaptation in order to increase energy flows. Larger cities invest more in maintaining and growing their dissipative

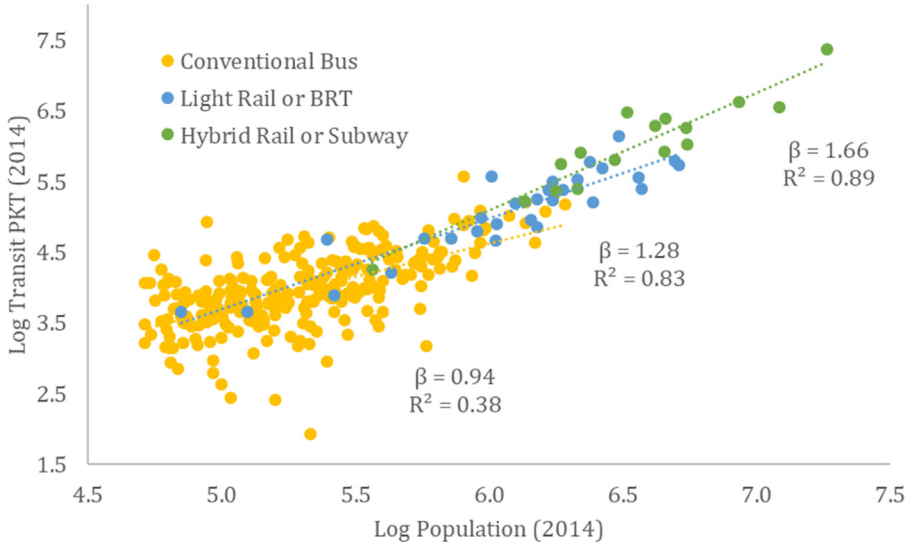

Fig. 5. Scaling relationships for Passenger-Kilometers-Traveled (PKT), for cities with Conventional Bus systems (yellow; $\beta = 0.94 \pm 0.15$, $R^2 = 0.38$, n = 253), Light Rail or Bus Rapid Transit (BRT) systems (blue; $\beta = 1.28 \pm 0.23$, $R^2 = 0.83$, n = 29), and Hybrid Rail or Subway systems (green; $\beta = 1.66 \pm 0.34$, $R^2 = 0.89$, n = 16). Data is for metropolitan statistical areas in the United States.
Source: [14].

structures, and in turn have more internal activity and energy inflow. When faced with problems such as increasing congestion costs, larger cities build higher-order infrastructure that makes their transportation systems more efficient, but also drives urban activity and growth. In rapidly urbanizing regions, this points to the need for strategies that can sustain potentially high energy use while reducing adverse impacts, such as local and global pollution.

Therefore, we can frame the contradiction between minimization of energy dissipation in natural systems and maximization of energy dissipation in cities in the following manner. Systems did not evolve to minimize energy dissipation; they evolved to maximize energy available to them—and natural systems have adapted more effectively than human systems. Natural systems have adapted to the constant energy source of the sun, water flow, etc. by optimizing feedbacks, recycling materials, and maintaining high quality storage. Their dissipation is minimized because their system is optimized for a constant energy flow. Human systems, in contrast, have focused their adaptation on the feedback mechanism that *increases* the flow of energy: burning fossil fuels, seeking fossil fuel reserves in more difficult-to-reach locations, nuclear energy, etc. In this sense, reducing individual energy consumption would in fact make more stored energy available to put towards the effort of increasing and using energy flows (i.e., growth).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix. Method

A.1. Scaling relationships

Empirical scaling relationships are determined based on the following general function:

$$Y = Y_0 N^{\beta}$$

where data fits were determined using ordinary least-squares in R. Adjustments were made for heteroskedasticity when present, as determined by a Breusch–Pagan test, using robust standard errors. Empirical scaling laws and figures were created with publicly available data for U.S. urbanized areas or metropolitan statistical areas.

A.2. Delineation of city boundaries

Empirical scaling relationships for population density for U.S. cities vary with the definition of the city boundary. The theoretical scaling exponent for population density, approximately 1/3, is rooted in Bettercourt's (2013) theoretical derivation of the sublinearity of area, particularly that the total power spent in transport processes to keep the city mixed is approximately equal to the total urban socioeconomic output. An important assumption in the model is that every person can reach every part of the city; in other words, the city has perfect mixing of populations. Bettencourt outlines that the cost of keeping the city mixed must be covered by each individual's budget, leading to the derivation for $A \propto N^{\alpha}$; $\alpha \approx 2/3$.

There are two geographical definitions for cities that are used by the U.S. Census Bureau: Metropolitan Statistical Area (MSA) and Urbanized Area (UA). The U.S. Office of Management & Budget [15] defines MSAs as the area surrounding a central urban area that represents the commuter shed; that is, it encompasses the area where at least 25 percent of residents commute to the central urban area for work. On the other hand, UAs are defined as "settled census tracts and blocks and adjacent densely settled territory that together contain at least 50,000 people" [15].

The scaling exponent for MSA population density is 0.51 ± 0.07, and the scaling exponent for UA population density is 0.12 ± 0.03. These two empirical values fall on either size of the theoretically predicted scaling relationship of 1/3 (i.e., 0.33), likely a reflection that neither UAs nor MSAs quite capture the perfect social mixing assumed by the model. People commuting into the city are missed in the definition of UAs, while the definition for MSAs includes large proportions of people who do not travel to the city center. In other words, in searching for the ideal real-world definition for a city that matches the definition in the model, UAs are not quite large enough, while MSAs are a bit too large.

References

[1] L. Bettencourt, D. Lobo, C. Helbing, G. Kuhnert, J. West, Growth, innovation, scaling, and the pace of life in cities, Proc. Natl. Acad. Sci. 104 (17) (2007) 7301–7306.
[2] L. Bettencourt, G. West, A unified theory of urban living, Nature 467 (7318) (2010) 912–913.
[3] L. Bettencourt, The origins of scaling in cities, Science 340 (6139) (2013) 1438–1441.
[4] I. Prigogine, Time, structure and fluctuations, Science 201 (4358) (1978) 777–785.
[5] E.D. Schneider, J.J. Kay, Life as a manifestation of the second law of thermodynamics, Math. Comput. Modelling 19 (6–8) (1994) 25–48.
[6] J.J. Kay, Ecosystems as self-organizing holarchic open systems: Narratives and the second law of thermodynamics, in: S.E. Jorgensen, F. Müller (Eds.), Handbook of Ecosystem Theories and Management, Lewis Publishers, Boca Raton, FL, 2000.
[7] D. Bristow, C. Kennedy, Why do cities grow? Insights from nonequilibrium thermodynamics at the urban and global scales, J. Ind. Ecol. 19 (2) (2015) 211–221.
[8] D.S.L. Cardwell, From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age, Cornell University Press, Ithaca, New York, 1971, p. 336.
[9] H.T. Odum, E.C. Odum, Energy Basis for Man and Nature, McGraw-Hill Book Company, New York, 1976.
[10] Texas A & M Transportation Institute and INRIX, Mobility scorecard, 2018, Available: https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015.pdf. (Accessed 15 February 2018).
[11] U.S. Department of Commerce, Bureau of Economic Analysis (BEA), Regional economic accounts – GDP by metropolitan area, 2016, Available: http://www.bea.gov/regional/index.htm. (Accessed 04 May 2016).
[12] Brookings Institution, Export monitor 2015, 2015, Available: https://www.brookings.edu/interactives/export-monitor-2015/ (Accessed 21 March 2018).
[13] U.S. Census Bureau, 2010 urban area to metropolitan and micropolitan statistical areas relationship file, 2015, Available: https://www.census.gov/geo/maps-data/data/ua_rel_download.html. (Accessed 03 May 2016).
[14] American Public Transportation Association, 2016 Public Transportation Fact Book, Appendix B: Operating Statistics and Rankings, 2017, Available: http://www.apta.com/resources/statistics/Documents/FactBook/2016-APTA-Fact-Book.pdf. (Accessed 21 March 2018).
[15] U.S. Office of Management and Budget (OMB), Standards for delineating metropolitan and micropolitan statistical areas; notice, 2010, Available: https://www.gpo.gov/fdsys/pkg/FR-2010-06-28/pdf/2010-15605.pdf. (Accessed 17 November 2018).