Protected-area boundaries as filters of plant invasions

Llewellyn C. Foxcroft[1,2,*], Vojtěch Jarošík[3,4], Petr Pyšek[4,3], David M. Richardson[2] & Mathieu Rouget[5]

[1]Conservation Services, South African National Parks, Skukuza 1350, South Africa

[2]Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

[3]Department of Ecology, Faculty of Sciences, Charles University in Prague, Viničná 7, CZ 128 44 Prague 2, Czech Republic

[4]Institute of Botany, Department of Invasion Ecology, Academy of Sciences of the Czech Republic, Průhonice, CZ 252 43, Czech Republic

[5]Department of Plant Science, University of Pretoria, Pretoria 0002, South Africa

[*]Author for correspondence: Llewellyn C. Foxcroft, Conservation Services, South African National Parks, Private Bag X402, Skukuza 1350, South Africa; llewellynf@sanparks.org; phone +27137354125, fax +27137354055.

Keywords

barriers to invasion, Kruger National Park, non-native invasive species, overland water flow, protected-area boundary

Running title:

Park boundaries and invasive plants

Manuscript information:

Word count: 3367

Number of references: **28**

Number of figures: **3** and tables: 0

Submission includes supplementary information.

Abstract

Human land uses surrounding protected areas provide propagules for colonization of these areas by non-native species, and corridors between protected-area networks and drainage systems of rivers provide pathways for long-distance dispersal of non-native species. However, the influence of protected-area boundaries on the colonization of protected areas by invasive non-native species is unknown. We drew on a spatially explicit data set of more than 27 000 non-native plant presence records for South Africa's Kruger National Park to examine the role of boundaries in preventing invasion of protected areas by non-native species. The number of records of non-native invasive plants declined rapidly beyond 1500 m inside the park, suggesting that the park boundary limited the spread of non-native plants from the surrounding landscape. The number of non-native invasive plants inside the park was a function of the amount of water runoff, density of major roads, and the presence of natural vegetation outside the park. Of the types of human-induced disturbance, only the density of major roads outside the protected area significantly increased the density of non-native plants. Our findings suggest that the probability of incursion of invasive plants into protected areas can be quantified reliably.

INTRODUCTION

Many anthropogenic threats to native species of plants and animals are not removed by establishing formal protected areas. Theoretically, the probability that formal conservation measures will meet their goals is improved by establishing networks of protected areas and by increasing connectivity through the creation of corridors (Foxcroft et al. 2007; Gaston et al. 2008). Such strategies, however, do little to prevent non-native invasive species from colonizing protected areas (Lambdon et al. 2008; Pyšek et al. 2008; Hulme et al. 2009). Some types of landscape features may exacerbate such colonizations; for example, river networks can facilitate colonization of non-native plants (Pyšek & Prach 1993; Margules & Sarkar 2007; Richardson et al. 2007). Protected-area management strategies that address colonization of non-native species generally focus on early detection and eradication, and action is applied only to species that are likely to have greatest negative effects on ecosystem functions. Establishment of buffer zones around protected areas is often included in these strategies. Although some researchers have addressed non-native species colonization at the interface between protected areas and human-dominated systems (Pyšek et al. 2003; Alston & Richardson 2006), no one has addressed the distance of non-native species' incursions into protected areas or explored what would constitute an effective and sustainable width of buffer to reduce incursions.

A data set on distributions of non-native plant species and land use in and around South Africa's Kruger National Park (KNP) allowed us to conduct a detailed analysis of the permeability of protected area boundaries. Kruger National Park in northeastern South Africa (Fig. 1), was founded in 1898, and covers an area of ~20,000 km^2. More than 370 non-native species have been recorded in the park (Foxcroft & Freitag-Ronaldson 2007). In response, KNP managers have initiated programs aimed at preventing and mitigating colonization of the park by non-native species (Foxcroft & Downey 2008; Koenig 2009), and detailed data on the distribution of these species have been collected over 4 years (Foxcroft et al. 2009).

We explored the extent to which park boundaries act as a barrier to the influx of non-native plant species from outside to inside the park. We also examined how particular characteristics of the surrounding landscape, such as land use and presence of features thought to promote dispersal of non-native plants, affect the number and location of records of non-native plants inside the park. Thus, we sought to identify factors that could be used to explain which areas more likely to be invaded, and how far invasions of non-native species extend into the park. Such insights are needed to explain invasions patterns and elucidate management options in this and other protected areas.

METHODS

Study area

We assessed the relation between the western and southern boundaries of KNP and colonization of the park by non-native invasive plants. We excluded the northern (Limpopo River) and eastern (border with Mozambique) boundaries from our analyses (Fig. 1). We based this delimitation on the assumption that propagules of non-native species arrive mainly from the western side of the KNP because all rivers flowing through the park flow from west to east and tourist entrance gates are primarily along the western and southern boundaries. Also, data from outside South Africa (Mozambique in the east and Zimbabwe in the north) are not as spatially comprehensive as data from South Africa. Additionally, the Limpopo River has an extensive drainage basin of which the KNP is only 4%; including this edge would thus distort the effects of water run-off on the potential colonization of non-native plants.

Data set

Approximately 120 field rangers collect data during their daily patrols with a hand-held personal computer and customized software (CyberTracker; MacFadyen 2005; Foxcroft et al. 2009). Rangers record data opportunistically as they move through an area. Apart from the presence of non-native plants, rangers also recorded observations of animals, carcasses, and tracks and water availability (Foxcroft et al. 2009), which we used as absence points. We considered a non-native species to be present at a given point if a ranger recorded its occurrence and absent if a ranger recorded a visit to the point but did not detect a plant, or recorded any other observations. Absence records were assumed to be accurate on the basis of the assumption that if a non-native plant was present at the same point as another observation, it would also have been recorded. We believe this assumption is justified because we considered only the most abundant and conspicuous non-native species that trained rangers could recognize reliably: erect pricklypear (*Opuntia stricta*), lantana (*Lantana camara*), triffid weed (*Chromolaena odorata*), and parthenium (*Parthenium hysterophorus*) (Foxcroft et al. 2009).

The data set is spatially explicit, covers all of KNP (Foxcroft et al. 2009), and includes >27,000 presence records and >2 million absence records collected between 2004 and 2007. Along the western and southern park boundaries we delineated 638 1-km-wide segments that extended into the park up to 1500 m (Fig. 1).

Non-native species

To characterize the occurrence of non-native species, we extracted numbers of their records, and to account for a possible effect of trends in sampling intensity by the park staff, also proportions of non-native records. To account for a possible effect of sampling intensity by rangers, we determined the perpendicular distances of all non-native plant records (hereafter non-native records), and the proportions of non-native records to all records, from the park boundary. For each segment we summed the distances within 100-m increments along a perpendicular line that ran from the boundary across the entire park (maximum distance was 52,000 m). Proportions of non-native records were calculated as the number of non-native records divided by the total number of non-native and absence records in the 100-m increments from the boundary up to a distance of 1700 m.

Environmental data

We examined the effects of several environmental variables on the presence of non-native plants in KNP. Inside the park, we summed the presence of roads, camps, gates, rivers, and the dominant type of vegetation within each segment along the boundary (Fig. 1). Outside the park, for sections corresponding to segments starting at the boundary and extending 1, 5, 10, and 50 km away we determined the density of major and secondary roads, land use, mean annual water runoff, presence of protected area, and primary productivity of vegetation (see Appendix S1 and S2).

Statistical analyses

We plotted the numbers and proportions of non-native records against distance from the boundary, and analyzed with Kendall's rank correlation whether the occurrence of non-native plants decreased monotonically as distance from the KNP boundary increased (Legendre & Legendre 1998). To determine whether there was a threshold distance at which the proportion of non-native records leveled off, we used a locally weighted scatterplot smoothing regression model (LOWESS) to assess a region of effective neutrality (Trexler & Travis 1993), defined as the distance from KNP boundary at which there was no relation between the distance and the number of non-native records (Cleveland 1979; Chambers et al. 1983).

We used a binary classification tree (Breiman et al. 1984; De'ath & Fabricius 2000) to analyze the presence and absence of non-native species in the segments as a function of environmental characteristics measured within and outside KNP. Classification trees sub-divide the dataset into homogenous groups based on a series of hierarchical splits. The classification accuracy of each split in the tree was expressed by its improvement value, corresponding to the overall misclassification rate at each node, with high scores of improvement values corresponding to splits of high quality. The quality of the chosen tree was evaluated as the overall misclassification rate by comparing the misclassification rate of the optimal tree with 50% misclassification rate of the null model (De'ath & Fabricius 2000), and using cross-validated samples (Steinberg & Colla 1995), based on values of sensitivity (the ability of the model to predict that a non-native species is present when it is) and specificity (the ability of the model to predict that a non-native species is not present when it is not) (Bourg et al. 2005).

The environmental variables identified by the optimal classification tree as explaining the greatest variance in number of non-native records were not collinear. Thus, we used them as explanatory variables in a logistic regression with binomial errors and logit link function (Quinn & Keough 2002) in which presence or absence of non-native species in each of the 638 segments was the response variable.

Details of these analyses are described in Appendix S3.

RESULTS

Inside the park, the number of non-native records decreased monotonically as distance from the KNP boundary increased (Fig. 2a,b) (Kendall's rank correlation $\tau = -0.53$; $z = -18.00$; $p<0.0001$ for number of non-native records). The proportion of non-native records also decreased monotonically as distance from the park boundary increased (Kendall's rank correlation $\tau = -0.71$; $z = -3.95$; $p = 0.0001$). The difference between non-native records and the proportion of records, up to 1700 m from the boundary was only 1.5% of the explained variance, whereas 93.6% of the explained variance was the same for the two measures (Appendix S4). This indicates that sampling bias associated with use of number rather than proportion of non-native records was negligible. At approximately 1500 m from the KNP boundary the number of non-native records became constant (Fig. 2b).

Mean annual water runoff (>6 million m^3/annum) from the watershed outside the park explained the greatest proportion of variance in non-native records in a given segment. Segments with less than mean runoff were more likely to have non-native species present only in areas with high road density within 10 km outside the park boundary (Fig. 3). The effects of water runoff and road density interacted in such a way that high run-off compensated for a low road density and vice versa, but high values of both factors at the same time did not result in an increased effect compared to a separate effect of either factor (Appendix S5). In a model in which mean annual runoff was replaced by a categorical measure of runoff (none [no rivers intersected the segment, low [2–10], medium [10–15], and high [15–26] million m^3 / quaternary watershed / annum), the number of non-native records in segments with no river present was associated again with the density of major roads within a 10-km radius outside the KNP boundary. If a segment contained a river, presence of non-native plants was unlikely only in segments where natural vegetation (possibly grazed by livestock) comprised over 90% of the land use in the 5-km radius outside the KNP boundary and where roads were absent within the park (Appendix S6).

DISCUSSION

Our results suggest that the park boundary limits the spread of non-native plants into the KNP. We quantified the number of records of non-native species, which we assumed is a surrogate measure of boundary permeability, on the basis of a few environmental variables measured outside and within the park. These features are related to factors that facilitate the spread of non-native species, such as the presence of water courses (Pyšek & Prach 1993; Richardson et al. 2007), human activity (von der Lippe & Kowarik 2007, Pyšek et al. 2010), and road construction. Our statistical models had high explanatory power and we believe they may be applicable to protected areas in general.

Of the 2 types of vectors of propagules of non-native plants, water runoff and road density, the association of water runoff with number of non-native plants was stronger. Park managers have little control over the upper reaches of the rivers that flow through KNP. But, it appears there is a quantifiable threshold value of water runoff from surrounding areas below which invasion of non-native plants is less likely. This knowledge could be applied to prioritization of measures to control colonization of areas by non-native plants, such as targeting particular riparian areas for removal of non-native plants. The effects of water and road density on dispersal of non-native plants appeared to be similar but not synergistic. The number of non-native records was explained by environmental variables up to a distance of 10 km from the boundary; therefore, data on variables that explain significant variance in presence of non-native invasive species can be collected relatively near the protected area.

These variables are associated with the spread of non-native plants into the park from surrounding landscapes. Within about 1500 m into the park from the boundary, the number of non-native plants decreased sharply. This distance beyond which a rapid decline in abundance of non-native plants occurs is likely to differ among protected areas, ecosystems, and vegetation types, but we believe such a threshold is likely to be present in all large protected areas. This does not mean park interiors cannot have high abundances of non-native plants in places. Land uses within a protected area can affect the spread of non-native plants (e.g., tourist camps and staff villages) and serve as sources of non-native species (Foxcroft 2001; Foxcroft et al. 2008). Further spread of non-native plants into KNP may be limited by placement of entrance gates where water runoff is low and the KNP border surrounded by natural areas. Additionally, because number of non-native plant records declined rapidly as distance into the park increased, we believe that a buffer zone of undisturbed vegetation surrounding the park could slow the spread of non-native plants.

Acknowledgments

We thank S. MacFadyen for the use of the data set collected through the CyberTracker programme and J. Pergl and Z. Sixtová for technical assistance. L.C.F thanks SANParks, the Department of Science and Technology – National Research Foundation, Centre of Excellence for Invasion Biology (C•I•B), and National Research Foundation Incentive Fund for support. P.P. and V.J. were supported by PRATIQUE (project KBBE-212459) of the 7th Framework Programme of the European Union and grant numbers AV0Z60050516 (AS CR), 0021620828, LC06073 (MSMT CR) and GA CR 206/09/0563. D.M.R acknowledges support from C•I•B, National Research Foundation, and Hans Sigrist Foundation, and P.P. from the Praemium Academiae award from the Academy of Sciences of the Czech Republic.

Supporting Information

A description of the environmental variables (Appendix S1) and vegetation type (Appendix S2), used to explain the occurrence of non-native plants in the KNP. Further details on the statistical analyses (Appendix S3), a comparison of number and proportion of presence records of non-native plants (Appendix S4), the logistic model explaining the probability of non-native plant presence in boundary segments (Appendix S5) and alternative model to optimal classification tree (Appendix S6) are available online. The authors are responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Alston, K. P., and D. M. Richardson. 2006. The roles of habitat features, disturbance, and distance from putative source populations in structuring alien plant invasions at the urban/wildland interface on the Cape Peninsula, South Africa. Biological Conservation **132:**83–198.

Bourg, N. A., W. J. McShea, and D. E. Gill. 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. Ecology **86:**2793–2804.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. 1984. Classification and regression trees. Wadsworth International Group, Belmont.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical methods for data analysis. Wadsworth International Group, Belmont, California.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of American Statistical Association **74:**829–836.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology **81:**3178–3192.

Foxcroft, L. C. 2001. A case study of human dimensions in invasion and control of alien plants in the personnel villages of Kruger National Park. Pages 127-134 in J. A. McNeely, editor. The great reshuffling: human dimensions of invasive alien species. International Union for Conservation of Nature, Gland, Switzerland.

Foxcroft, L. C., and P. O. Downey. 2008. Protecting biodiversity by managing alien plants in national parks: perspectives from South Africa and Australia. Pages 387–403 in B. Tokarska-Guzik, J. H. Brock, G. Brundu, L. Child, C. C. Daehler, and P. Pyšek, editors. Plant invasions: human perception, ecological impacts and management. Backhuys Publishers, Leiden, Netherlands.

Foxcroft, L. C., and S. Freitag-Ronaldson. 2007. Seven decades of institutional learning: managing alien plant invasions in the Kruger National Park, South Africa. Oryx **41:**160–167.

Foxcroft, L. C., D. M. Richardson, M. Rouget, and S. MacFadyen. 2009. Patterns of alien plant distribution at multiple spatial scales in a large national park: implications for ecology, management and monitoring. Diversity and Distributions **15:**367–378.

Foxcroft, L. C., D. M. Richardson, and J. R. U. Wilson. 2008. Ornamental plants as invasive aliens: problems and solutions in Kruger National Park, South Africa. Environmental Management **41:**32–51.

Foxcroft, L. C., M. Rouget, and D. M. Richardson. 2007. Risk assessment of riparian plant invasions into protected areas. Conservation Biology **21:**412–421.

Gaston, K. J., S. F. Jackson, L. Cantú-Salazar, and G. Cruz-Piñón. 2008. The ecological performance of protected areas. Annual Review of Ecology, Evolution, and Systematics **39:**93–113.

Hosmer, D. W., and S. Lemeshow. 1989. Applied logistic regression. Wiley, New York.

Hulme, P., P. Pyšek, W. Nentwig, and M. Vilà. 2009. Will threat of biological invasions unite the European Union? Science **324:**40–41.

Koenig, R. 2009. Unleashing an army to repair alien-ravaged ecosystems. Science **325:**562–563.

Lambdon, P. W. et al. 2008. Alien flora of Europe: species diversity, temporal trends, geographical patterns and research needs. Preslia **80:**101–149.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.

MacFadyen, S. 2005. The Kruger National Park CyberTracker Program- electronic ranger diaries. South African National Parks, Skukuza, South Africa. Available from www.sanparks.org (accessed December 2009).

Margules, C. R., and S. Sarkar. 2007. Systematic conservation planning. Cambridge University Press, Cambridge, United Kingdom.

Pyšek P., et al. 2010. Disentangling the role of environmental and human pressures on biological invasions. Proceedings of the National Academy of Sciences of the United States of America 107: 12157–12162.

Pyšek, P., V. Jarošík, and T. Kučera. 2003. Inclusion of native and alien species in temperate nature reserves: an historical study from Central Europe. Conservation Biology **17:**1414–1424.

Pyšek, P., and K. Prach. 1993. Plant invasions and the role of riparian habitats: a comparison of four species alien to central Europe. Journal of Biogeography **20:**413–420.

Pyšek, P., D. M. Richardson, J. Pergl, V. Jarošík, Z. Sixtová, and E. Weber. 2008. Geographical and taxonomic biases in invasion ecology. Trends in Ecology & Evolution **23:**237–244.

Quinn, G. P., and M. J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, United Kingdom.

Richardson, D. M., P. M. Holmes, K. J. Esler, S. M. Galatowitsch, J. C. Stromberg, S. P. Kirkman, P. Pyšek, and R. J. Hobbs. 2007. Riparian vegetation: degradation, alien plant invasions, and restoration prospects. Diversity and Distributions **13:**126–139.

Steinberg, G., and P. Colla. 1995. CART: Tree-structured Non-parametric Data Analysis. Salford Systems, San Diego.

Trexler, J. C., and J. Travis. 1993. Nontraditional regression analyses. Ecology **74:**1629–1637.

von der Lippe, M., and I. Kowarik. 2007. Long-distance dispersal of plants by vehicles as a driver of plant invasions. Conservation Biology **21:**986–996.

Figure Captions

Figure 1. Mean annual river run off in Kruger National Park (KNP); pattern of non-native plant records in relation to KNP boundary and 1 × 1.5 km segments (lower left inset); and location of KNP within South Africa (upper left inset).

Figure 2. Number of presence records of non-native plants relative to distance from the boundary of Kruger National Park toward the interior of the park: (a) to 52,000 m and (b) to 3,000 m. Fitted curves are LOWESS regression models on square root number of records (back-transformed for visualization) with no patterns of residuals: In (a) span of smoothing = 0.1; equivalent number of parameters in curvilinear regression ENP=29.8; explained variance $r^2 = 0.82$; model chosen by deletion test (F=1.16; ENP = 313.2, 29.8; p=0.13) against starting model with span = 0.01 and $r^2 =$ 0.94. In (b) span of smoothing = 0.5; ENP = 6.4; $r^2 = 0.96$; model chosen by deletion test (F=2.37; ENP=15.1, 6.4; p=0.07) against starting model with span = 0.2 and $r^2 = 0.99$. The spans of smoothing describe local sensitivity in iterations of LOWESS models (see Appendix S3 for details).

Figure 3. Results of classification tree analysis of the binary probability of non-native species presence in Kruger National Park (KNP). Probability of presence was determined on the basis of water runoff in the park (mean annual runoff from the surrounding watershed; million m^3) and road density adjacent to the park (density of major roads within 10 km radius outside the KNP boundary) (%, percentage of cases for each class; bars, representation of percentage of absent [grey] and present [black]). Except for the root node (undivided data) at the top, there is splitting variable name and split criterion above each node. Vertical depth of each node is proportional to its improvement value. Overall misclassification rate of the optimal tree is 14.0%, sensitivity 0.92 and specificity 0.81.

(summarized for sections adjacent to each 1 km wide transect in a 1, 5, 10 and 50 km radius outside the park boundary).

[a]Major roads are defined as the main tourist tar roads in KNP

[b]Natural areas are untransformed landscapes (although probably grazed by livestock), cultivated areas are agricultural (crop) lands, plantations are commercial plantation forests, urban areas include towns and informal / rural settlements, degraded areas are areas transformed by erosion (gullies and bare soil), loss of plant cover, and other disturbances

[c]Refers to quaternary watersheds; Quaternary watersheds are nested subdivisions of primary, secondary and tertiary watersheds, where primary watersheds refer to the drainage areas of major rivers in the South African river classification system (van Wilgen *et al.* 2007). Run-off was only given for those segments for which a main river intersected it.

[d]Normalized Difference Vegetation Index; is a measure of the amount of green vegetation i.e. photosynthetically active material, and is used as a proxy for above-ground net primary production.

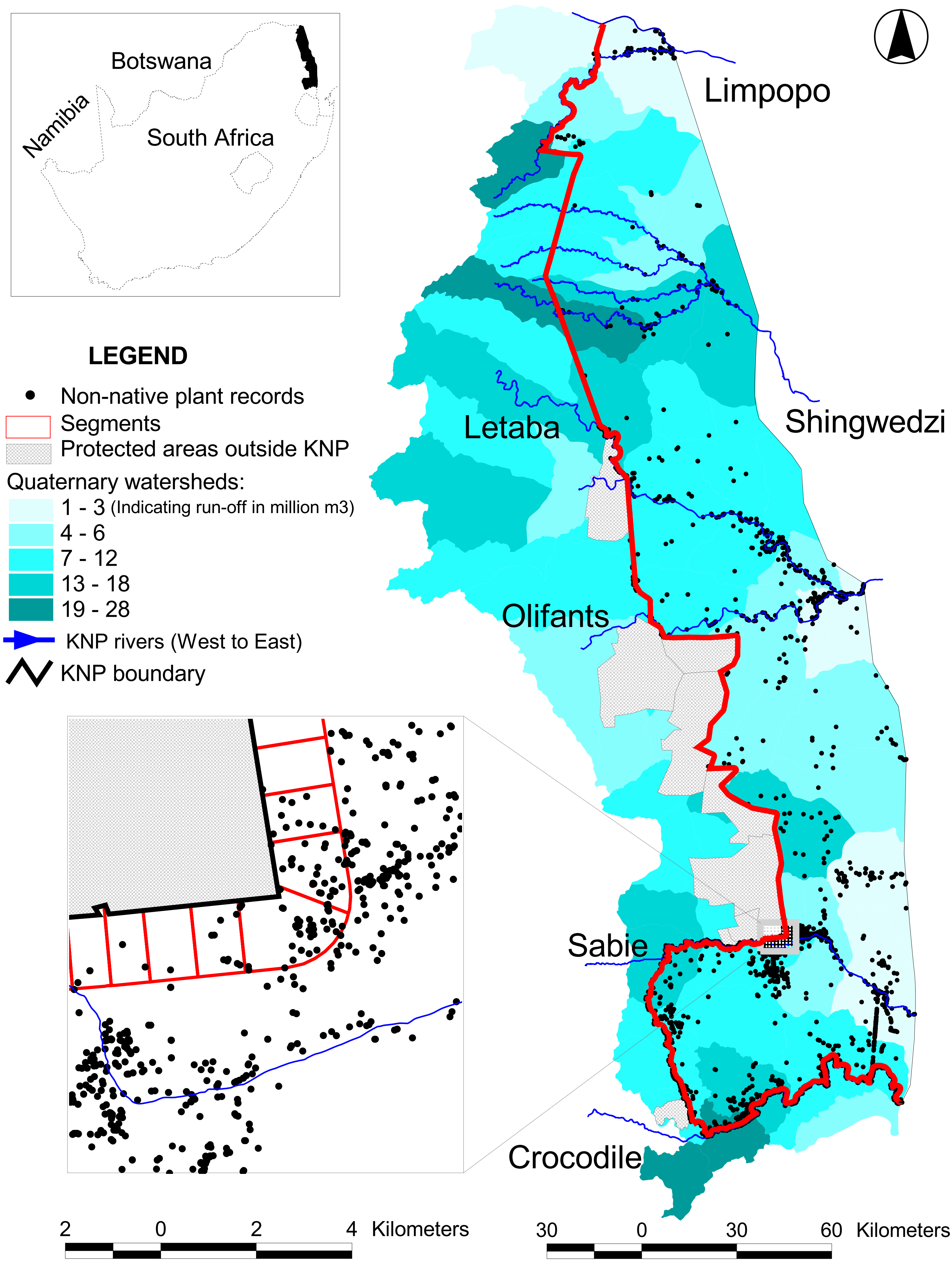
[e]Main rivers are defined as the seven main river systems (Limpopo, Luvuvhu, Shingwedzi, Letaba, Olifants, Sabie, Crocodile Rivers; Fig. 1) flowing into and through the KNP from west to east

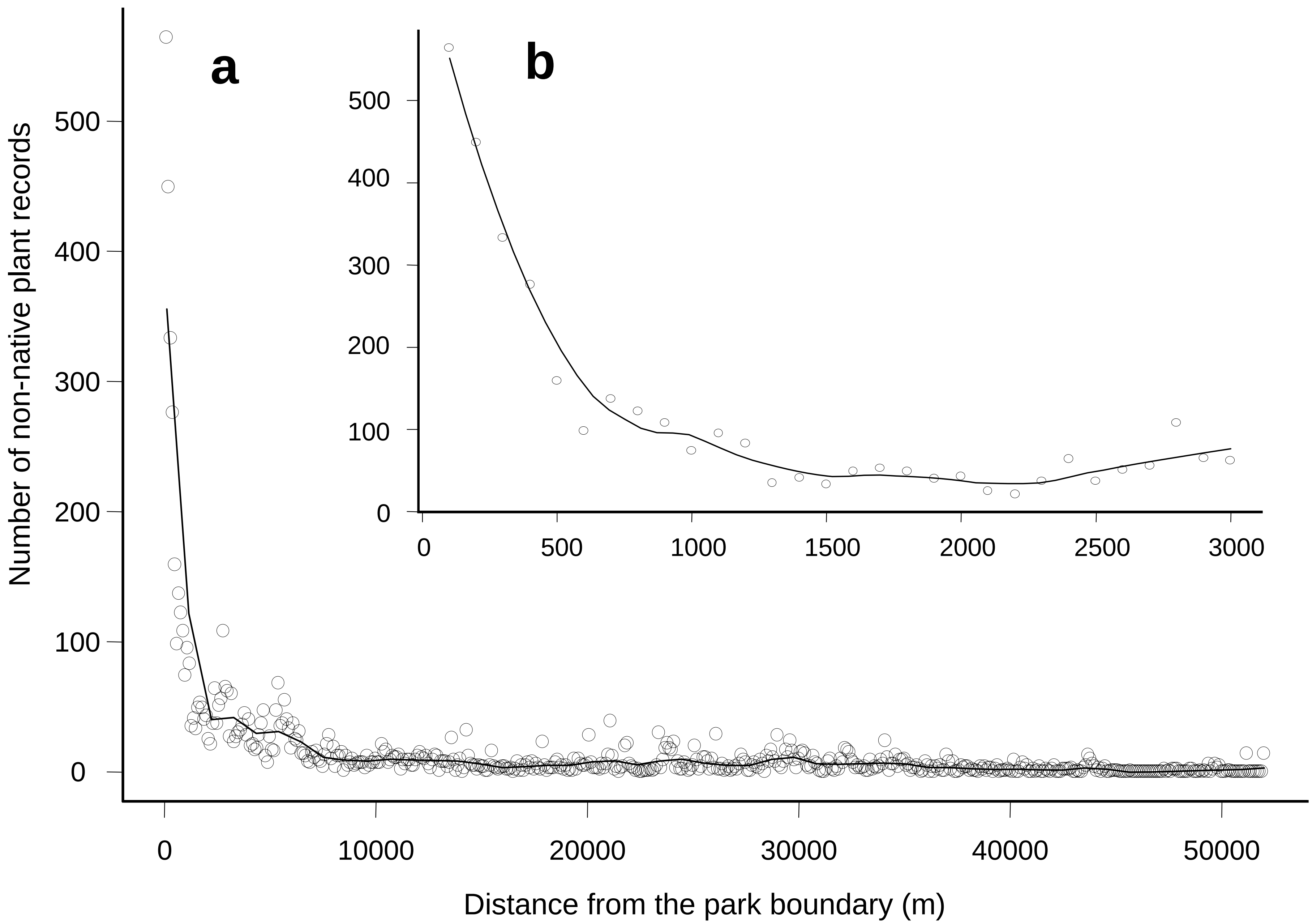
[f]See Appendix 2.

References

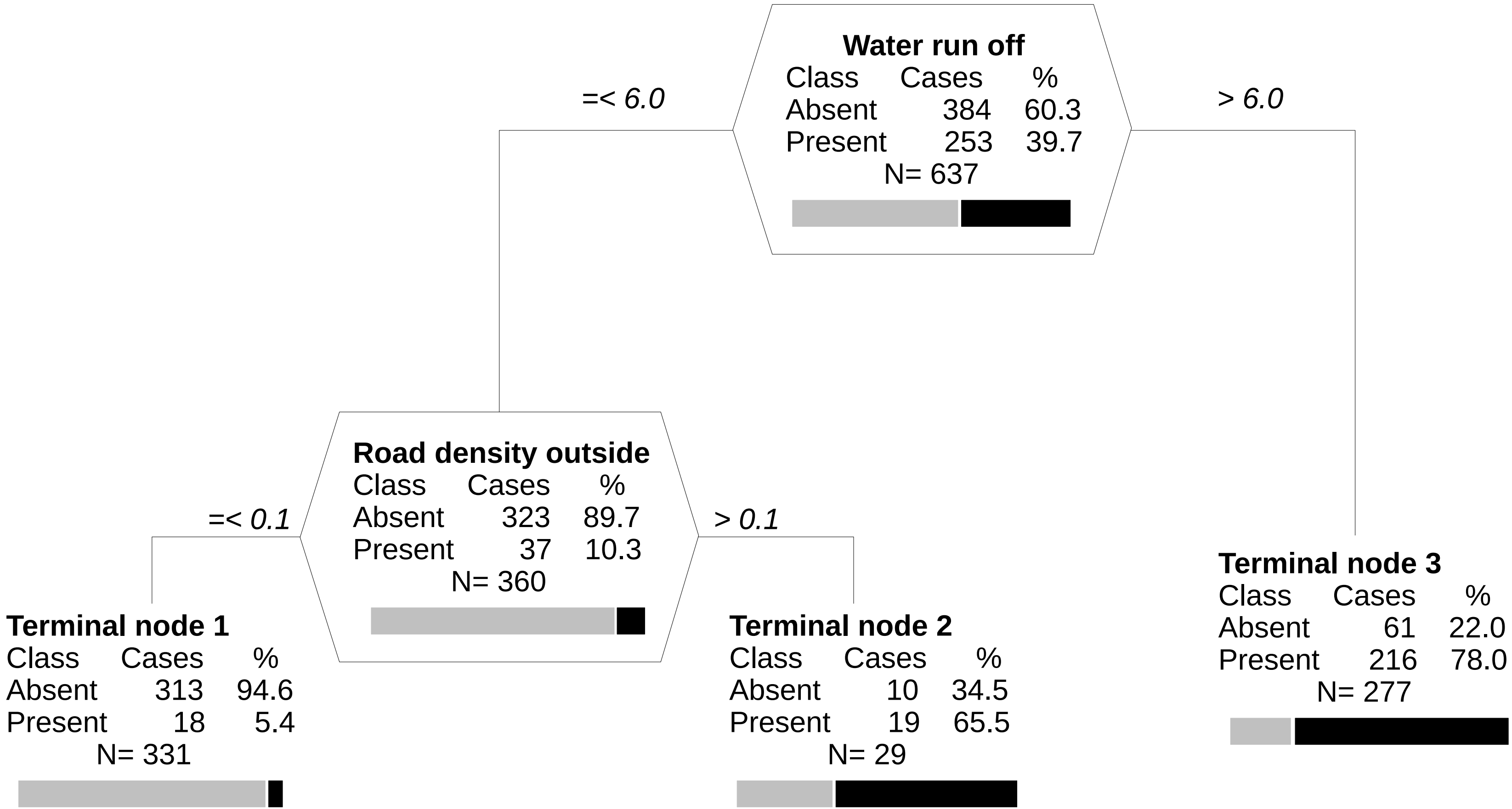
Gertenbach, W. P. D. 1983. Landscapes of the Kruger National Park. Koedoe **26:**9-121.

van Wilgen, B. W., J. L. Nel, and M. Rouget. 2007. Invasive alien plants and South African rivers: a proposed approach to the prioritization of control operations. Freshwater Biology **52:**711–723.


Botswana
Namibia
South Africa
Limpopo
Shingwedzi
Letaba
Olifants
Sabie
Crocodile
LEGEND
Non-native plant records
Segments
Protected areas outside KNP
Quaternary watersheds:
1 - 3 (Indicating run-off in million m3)
4 - 6
7 - 12
13 - 18
19 - 28
KNP rivers (West to East)
KNP boundary
2 0 2 4 Kilometers
30 0 30 60 Kilometers



a
b
Number of non-native plant records
Distance from the park boundary (m)
0
100
200
300
400
500
0
10000
20000
30000
40000
50000
0
500
1000
1500
2000
2500
3000


Water run off
Class Cases %
Absent 384 60.3
Present 253 39.7
N= 637
=< 6.0
> 6.0
Road density outside
Class Cases %
Absent 323 89.7
Present 37 10.3
N= 360
=< 0.1
> 0.1
Terminal node 1
Class Cases %
Absent 313 94.6
Present 18 5.4
N= 331
Terminal node 2
Class Cases %
Absent 10 34.5
Present 19 65.5
N= 29
Terminal node 3
Class Cases %
Absent 61 22.0
Present 216 78.0
N= 277

Supporting Information: Appendix 1

Table. Environmental variables used to explain the occurrence of non-native plants in the Kruger National Park (KNP).

	Category	Variable	Variable type
Outside KNP	[a]Traffic	Density of major national roads within 1, 5, 10, 50 km of boundary [km/km^2]	Continuous
		Density of all (major, secondary and gravel) roads 1, 5, 10, 50 km [km/km^2]	Continuous
	[b]Land use	% natural areas in 1, 5, 10, 50 km	Continuous
		% cultivated areas in 1, 5, 10, 50 km	Continuous
		% plantations in 10, 50 km	Continuous
		% urban areas in 1, 5, 10, 50 km	Continuous
		% degraded areas in 1, 5, 10, 50 km	Continuous
	Protected areas	Protected areas adjacent KNP	Binary (present/absent)
	[c]Run-off from quaternary watershed	Mean annual runoff [million m^3 / quaternary watershed / annum]	Continuous
		River runoff category [none, low, medium, high]. (*None* indicates that no main rivers intersected the segment)	Categorical
	Vegetation productivity	[d]NDVI mean value	Continuous
Inside KNP	Human activities	[a]Presence of major roads	Binary (present/absent)
		Presence of roads	Binary (present/absent)
		Presence of camps	Binary (present/absent)
		Presence of gates	Binary (present/absent)
	[e]Rivers	Presence of main river	Binary (present/absent)
		Presence of rivers	Binary (present/absent)
	[f]Vegetation type	Landscape unit (Gertenbach 1983)	Categorical

Notes: Data was extracted for areas inside the park (summarized in 1 × 1.5 km segments on transects from the edge of the park towards its interior; Fig. 1), and outside the park

Supporting Information: Appendix 2

Table. Landscape units of Kruger National Park (KNP) used as an explanatory variable for non-native plant records in KNP.

Landscape ID	Landscape*	Description
1	Lowveld Sour Bushveld of Pretoriuskop	Granitic plains with *Terminalia sericea* tree savanna
2	Malelane Mountain Bushveld	Granitic mountains with *Combretum apiculatum* bush savanna
4	Thickets of the Sabie & Crocodile Rivers	Granitic lowlands with *Acacia grandicornuta* tree savanna
5	Mixed *Combretum* / *Terminalia sericea* woodland	Granite plains with *Combretum zeyheri* or *Combretum apiculatum* bush savanna
6	Combretum / Colophospermum mopane woodland of Timbavati	Granite plains with *Colophospermum mopane* bush or tree savanna
7	Olifants River Rugged Veld	Granite plains with *Colophospermum mopane* bush or tree savanna
8	Phalaborwa Sandveld	Granitic plains with *Colophospermum mopane* bush savanna
9	*Colophospermum mopane* woodland / savanna on basic soil	Metalava with *Colophospermum mopane* tree savanna
10	Letaba River Rugged Veld	Metalava with *Colophospermum mopane* tree savanna
11	Tsende Sandveld	Granitic plains with *Colophospermum mopane* bush savanna
12	*Colophospermum mopane* / *Acacia nigrescens* savanna	Metalava plains with *Colophospermum mopane* tree savanna or Andesitic plains with *Combretum collinum* shrub savanna
13	*Acacia welwitschii* thickets on Karoo sediments	Karoo sediment plains with *Acacia welwitschii* tree savanna or with *Terminalia sericea* bush savanna
16	Punda Maria Sandveld on Cave Sandstone	Clarens sandstone hills with *Terminalia sericea* bush savanna or Soutpansberg group

		mountains with *Burkea africana* tree savanna
17	*Sclerocarya birrea* subspecies *caffra* /*Acacia nigrescens* savanna	Basaltic or gabbroic plains with *Sclerocarya birrea* tree savanna or *Acacia nigrescens* bush or shrub savanna
19	Thornveld on gabbro	Basaltic or gabbroic plains with *Sclerocarya birrea* tree savanna or *Acacia nigrescens* bush or shrub savanna
24	*Colophospermum mopane* shrubveld on gabbro	Basaltic or gabbroic plains with *Acacia nigrescens* bush savanna or *Colophospermum mopane* bush or shrub savanna
25	*Adansonia digitata* / *Colophospermum mopane* Rugged Veld	Basaltic or calcitic plains with *Colophospermum mopane* shrub savanna
26	*Colophospermum mopane* shrubveld on calcrete	Basaltic or calcitic plains with *Colophospermum mopane* shrub savanna
28	Limpopo / Luvuvhu Floodplains	Alluvial plains with *Faidherbia albida* or *Salvadora angustifolia* tree savanna
29	Lebombo South	Basaltic plains or rhyolite mountains with *Combretum apiculatum* or *Colophospermum* mopane bush savanna
33	*Pterocarpus rotundifolius* / *Combretum collinum* woodland	Metalava plains with *Colophospermum mopane* tree savanna or Andesitic plains with *Combretum collinum* shrub savanna
34	Punda Maria Sandveld on Waterberg sandstone	Clarens sandstone hills with *Terminalia sericea* bush savanna or Soutpansberg group mountains with *Burkea africana* tree savanna

*Landscapes are defined as “An area with a specific geomorphology, macroclimate, soil and vegetation pattern, and associated fauna” (Gertenbach 1983).

Reference

Gertenbach, W. P. D. 1983. Landscapes of the Kruger National Park. Koedoe **26:**9–121.

Supporting Information: Appendix 3

Further details on statistical analysis.

The numbers of non-native records could be biased by sampling intensity, for example, the park staff might have been collecting records in the surroundings of the boundary more intensively than in park interior. To test if this was the case, the numbers and proportions of non-native records were compared for their trends with increasing distance from park boundary up to 1700 m, a distance for which both measures were available, by using curvilinear polynomial regression (Sokal & Rohlf 1995). Before this analysis, numbers were square rooted and proportions angular transformed to normalize the data (Sokal & Rohlf 1995) and standardized to zero mean and variance one, separately for each monitoring method, to make the two measures comparable. Trends were then measured by ANCOVA with monitoring method as a two level factor and powers of distant classes as covariates. The powers were added sequentially until the addition caused insignificant reduction in explained variance. The similarity between trends was then assessed by testing significance of interaction between monitoring method and distance classes in deletion tests. Explained variance of individual models was compared based on r^2_{adj}, which takes into account number of model parameters (e.g. Quinn & Keough 2002).

To reveal if there was a threshold distance from which the proportions of non-native records level off, we assessed a region of effective neutrality (Trexler & Travis 1993) by iterations of locally weighted scatterplot smoothing (LOWESS) regression model (Cleveland 1985; Chambers et al. 1983) on square root transformed non-native plant numbers. The iterations differed in local sensitivity expressed as a span of smoothing between 0–1, where span = 1 corresponds to standard linear regression, with complexity described as equivalent number of parameters (ENP) in curvilinear regression (Cleveland 1993). We started with a small value of span (low smoothing and high proportion of explained variance, Trexler & Travis 1993) and increased the span slowly to the point at which the LOWESS model significantly ($p<0.05$) differed in ANOVA deletion test from the starting model; a slightly smaller value of span, which did not differ significantly from the starting model, was chosen for final interpretation (S-PLUS 4 1997). Residuals of all models were checked to verify whether they do not show any pattern (Cleveland 1985). Calculations were done in S-PLUS v. 8.1.1 (TIBCO Software).

The permeability of park edge to invasions was analyzed by dividing segments into those with non-native species present and absent, therefore the response variable was yes/no, and environmental characteristics measured within the KNP and outside the park were explanatory variables. Proportional explanatory variables were angular transformed to normalize the data, continuous variables checked for normality and standardized to zero mean and unit variance to achieve their comparable influence. Using the standardized values, collinearity was checked by calculating tolerance values among all continuous explanatory variables. Tolerances less than 0.1 were considered as values which can prevent a reliable estimate of linear model parameters (Quinn & Keough 2002). These values appeared for 10 of the 28 examined variables outside the KNP boundary, and regularly repeated for some of the variables also when collinearity checked separately at 1, 5, 10 and 50 km radius outside the KNP boundary. Consequently, instead of generalized linear models (GLMs) which would give unreliable estimates, the question what factors influence the permeability of the edge was first examined by binary classification trees (Breiman et al. 1984), a non-parametric statistical method for which collinearity of the explanatory variables cannot prevent reliable estimates. The use of the trees appeared preferable to GLMs also because some of the explanatory variables were skewed to the left, and no transformation was able to normalize the data for their proper use in GLMs. In regression trees, non-normal distribution cannot prevent reliable estimates because the trees are invariant to monotonic transformations of explanatory variables, and the explanatory variables thus do not need any transformation before analyses (De'ath & Fabricius 2000).

The classification trees were constructed in CART v. 6.0 (Breiman et al. 1984; Steinberg & Colla 1995) by binary recursive partitioning, using the default "Gini" impurity measure as the splitting index. To determine the optimal tree, a sequence of nested trees of decreasing size, each of them being the best of all trees of its size, were constructed, and their resubstitution relative errors estimated. Ten-fold cross-validation was used to obtain estimates of cross-validated relative errors of these trees. These estimates were then plotted against tree size, and the optimal tree chosen both based on the minimum cost tree rule, which minimizes the cross validated error (the default setting in CART v 6.0; Steinberg & Colla 1995), and based on the (ii) 1–SE rule, which minimizes cross-validated error within one standard error of the minimum (Breiman et al. 1984). A series of 50 cross-validations were run, and the modal (most likely) single optimal tree chosen for description (De'ath & Fabricius 2000).

The quality of the chosen tree was evaluated as the overall misclassification rate by comparing the misclassification rate of the optimal tree with misclassification rate of the null model (De'ath & Fabricius 2000), and using cross-validated samples (Steinberg & Colla 1997), based on values of sensitivity (i.e. true positive rate, defined as proportion of observations correctly identified as suitable), specificity (true negative rate) (Bourg et al. 2005) and misclassification rate for the presence/absence of a non-native plant in a segment (Steinberg & Colla 1995). Because high categorical explanatory variables have higher splitting power than continuous variables, to prevent the high categorical explanatory variable landscape (22 categories) to have inherent advantage over continuous variables, penalization rules for high category variables (Steinberg & Colla 1997) were applied.

The environmental characteristics chosen by the optimal classification tree as the best predictors of the permeability of park edge lacked collinearity and could be used as explanatory variables in logistic regression (Chytrý et al. 2008; Hejda et al. 2009; Pyšek et al. 2009) with binomial errors and logit link function (Quinn & Keough 2002), in which non-native species presence or absence in individual segments was the response variable. Overall significance of the logistic regression and significance of individual parameters and their interaction were evaluated by deletion tests based on G^2 statistic, i.e. as the deviance of the maximum likelihood model (Crawley 2002). An overall goodness of fit of the model was assessed with the Hosmer and Lemeshow (1989) test and a determined lack of fit removed by assessing which segments had the largest influence on the model. It was done by assessment of $\Delta\beta$, the analogue of Cook's statistic in logistic regression (Hosmer & Lemeshow 1989), for the individual segments. Segments with the largest $\Delta\beta$ were sorted in descending order and weighted out of the analysis one after another (Gilchrist & Green 1994; Křivánek et al. 2006) until the value of Hosmer and Lemeshow test (1989) indicated no evidence of the lack of fit. The quality of the final model was expressed as r^2_L, i.e. as an analogue of the explained variance of the ordinary regression model (Quinn & Keough 2002; Menard 2000) and as a percentage of correctly classified values for the presence/absence of a non-native plant in a segment. To verify whether the neighboring segments do not exhibit more similar records than expected by chance (e.g. Legendre 1993; Lichstein 2002), spatial autocorrelations of the individual segments were calculated as autocorrelation coefficients (Legendre & Fortin 1989) on Pearson's standardized residuals of the logistic regression. Insignificant coefficients of the residuals indicated no spatial autocorrelations among the segments (Rangel et al. 2006; Dormann et al. 2007). Calculations were done in S-PLUS v. 8.1.1 (TIBCO Software) and SPSS v. 17 (SPSS Inc.).

References

Bourg, N. A., W. J. McShea, and D. E. Gill. 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. Ecology **86:**2793–2804.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. 1984. Classification and Regression Trees. Wadsworth International Group, California.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical Methods for Data Analysis. Wadsworth International Group, Belmont, California.

Chytrý, M., V. Jarošík, P. Pyšek, O. Hájek, I. Knollová, L. Tichý, and J. Danihelka. 2008. Separating habitat invasibility by alien plants from the actual level of invasion. Ecology **89:**1541–1553.

Cleveland, W. S. 1985. The Elements of Graphing Data. Monterey, Wadsworth, Californnia.

Cleveland, W. S., E. Grosse, and W. M. Shyu. 1993. Local Regression Models. Pages 309–376 in S. J. M. Chambers, and T. J. Hastie, editors. Statistical Models in S. Chapman & Hall, New York.

Crawley, M. J. 2002. Statistical Computing: an Introduction to Data Analysis Using S-Plus. Wiley, New York.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology **81:**3178–3192.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, et al. 2007. Incorporating spatial autocorrelation in the analysis of ecological species distribution data: a user's guide. Ecography **30:**609–628.

Gilchrist, R., and P. Green 1994. The theory of generalized linear models. Pages 259–305 in B. Francis, M. Green, and C. Paine, editors. The GLIM System Release 4 Manual. Clarendon Press, Oxford.

Hejda, M., P. Pyšek, and V. Jarošík. 2009. Impact of invasive plants on the species richness, diversity and composition of invaded communities. Journal of Ecology **97:**393–403.

Hosmer, D. W., and S. Lemeshow. 1989. Applied Logistic Regression. Wiley, New York.

Křivánek, M., P. Pyšek, and V. Jarošík. 2006. Planting history and propagule pressure as predictors of invasion by woody species in a temperate region. Conservation Biology **20:**1487–1498.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74:**1659–1673.

Legendre, P., and M.-J., Fortin. 1989. Spatial pattern and ecological analysis. Vegetatio **80:**107–138.

Lichstein, J. W., T. R. Simons, S. A. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs **72:**445–463.

Menard, S. 2000. Coefficient of determination for multiple logistic regression analysis. The American Statistician **54:**17–24.

Pyšek, P., M. Křivánek, and V. Jarošík. 2009. Planting intensity, residence time, and species traits determine invasion success of alien woody species. Ecology **90:**2734–2744.

Quinn, G. P., and M. J. Keough. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge.

Rangel, T. F. L. V. B., J. A. F. Diniz-Filho, and L. M. Bini. 2006. Spatial patterns of terrestrial vertebrate species richness in the Brazilian Cerrado. Global Ecology and Biogeography **15:**321–327.

S-PLUS 4. 1997. Guide to Statistics. MathSoft, Seattle.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Freeman, San Francisco.

Steinberg, G., and P. Colla. 1995. CART: Tree-structured Non-parametric Data Analysis. Salford Systems, San Diego.

Steinberg, G., and P. Colla. 1997. CART: Tree-structured Non-parametric Data Analysis. Salford Systems, San Diego.

Trexler, J. C., and J. Travis. 1993. Nontraditional regression analyses. Ecology **74:**1629–1637.

Supporting Information: Appendix 4

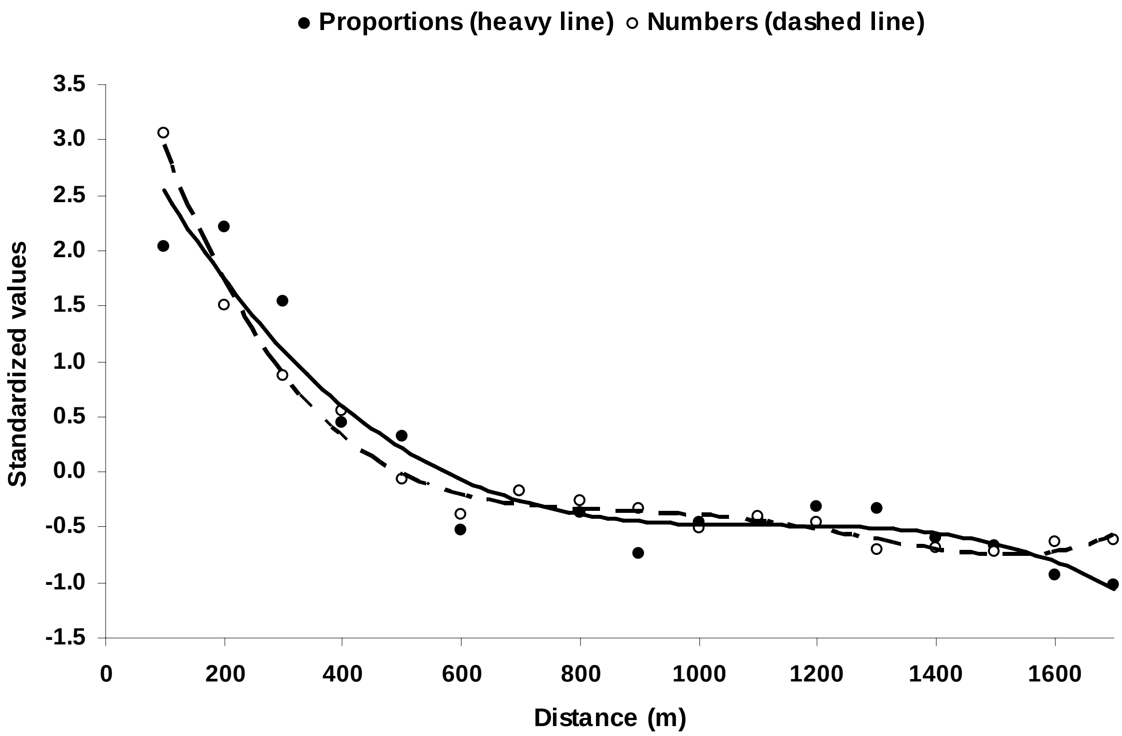


Figure. Comparison of non-native plant record trends in 100 m distance classes from the KNP boundary for two monitoring methods; non-native numbers and their proportional representation. The trends differ significantly only in the fourth power of polynomial function (deletion test on interaction between monitoring method and polynomial function which includes the fourth power: $F = 3.67$; $df = 4, 28$; $p = 0.018$; r^2_{adj} for model with the fourth power polynomial function = 0.951) but the differences are insignificant for polynomial function with only powers up to the third degree (deletion test on interaction between monitoring method and polynomial function up to the cubic power: $F = 0.46$; $df = 3, 29$; $p = 0.71$; r^2_{adj} for model with the third power polynomial function = 0.936); the fourth power appeared significant for the numbers (0.99 ± 0.31 [value ± standard error]; $t = 3.23$; $df = 24$; $p = 0.0035$), but insignificant (-0.55 ± 0.31; $t = 1.80$; $p = 0.084$) for the proportions. The minimal adequate models with all parameters significant, where x and their powers measure the distance from the boundary, are: proportions = $-3.34x + 1.63x^2 - 1.05x^3$ ($F = 59.95$; $df = 3, 13$; $p < 0.0001$; $r^2adj = 0.917$); numbers = $-3.15x + 2.01x^2 - 1.14x^3 + 0.70x^4$ ($F = 233.5$; $df = 4, 12$; $p < 0.0001$; $r^2adj = 0.983$); intercepts are zero due to standardizations.

Supporting Information: Appendix 5

Table. Summary and diagnostic information on the optimal classification tree for the probability of presence/absence (yes/no in the terminal node information below) of a non-native plant in Kruger National Park boundary segments, for Fig. 3.

Terminal node information. The first row for each terminal node shows the terminal node number (Node), number of cases at the node (N), fraction of the data that end up in the node (Prob), fraction of records at the node misclassified (Cost), and the dominant splitting category of the response variable for the node (Class). The following rows show distribution of the categories of the response variable at the node (N = numbers, Prob = probabilities of the numbers, Class = category).

```
=========================
TERMINAL NODE INFORMATION
=========================

 Node            N   Prob   Cost  Class
------------------------------------------
    1          331 0.4431 0.0803  NO

               313 0.9197         NO
                18 0.0803         YES

    2           29 0.0506 0.2575  YES

                10 0.2575         NO
                19 0.7425         YES

    3          277 0.5063 0.1569  YES

                61 0.1569         NO
               216 0.8431         YES
```

Misclassification by class. Category of the response variable (Class), probabilities of finding a member of each class in the whole sample (Prior Prob.), number of cases in each class (Count), number misclassified (Misclass), and rate of misclassification for each class (Cost). Data set used to grow the tree, i.e. the learning sample, shows the results when these data are dropped down the chosen tree. Data in brackets are based on cross validation, i.e. the best estimates of the misclassification that would occur if the classification tree were to be applied to new data, assuming that the new data were drawn from the same distribution as the learning data.

```
=========================
MISCLASSIFICATION BY CLASS
=========================

(Cross Validation)
              Prior
 Class         Prob         Count     Misclass         Cost
------------------------------------------------------------
NO          0.50000           384           71      0.18490
                             (384           73      0.19010)
YES         0.50000           253           18      0.07115
                             (253           19      0.07510)
------------------------------------------------------------
 Total      1.00000           637           89
                             (637           92)
```

Prediction success classification and probability tables. Cases of the classification and probability tables appearing on the diagonals of the matrices correspond to correct classification, while off-diagonal entries represent misclassification; the tables make it easy to see where the misclassifications are occurring. The learning sample tables are generated from the application of the input data to the selected optimal tree, and the cross-validation tables constitute the best estimate of the results that would occur if the tree were applied to fresh data. Pred. Tot. = predicted total in each column or class; Correct = percentage classified correctly; Tot. correct = fraction of cases classified correctly overall; Success ind. = fraction by which the percent correct exceeds what would be expected on the basis of chance alone.

```
=====================================
Cross Validation Classification Table
=====================================

Actual       Predicted Class                   Actual
Class                     NO          YES       Total
----------------------------------------------------
NO                    311.00        73.00      384.00
YES                    19.00       234.00      253.00
----------------------------------------------------
Pred. Tot.            330.00       307.00      637.00
Correct              0.80990      0.92490
Success Ind.         0.20707      0.52773
Tot. Correct         0.85557
-----------------------------------------------------

=================================================
Cross Validation Classification Probability Table
=================================================

Actual       Predicted Class                   Actual
Class                     NO          YES       Total
----------------------------------------------------
NO                   0.80990      0.19010     1.00000
YES                  0.07510      0.92490     1.00000
----------------------------------------------------

====================================
Learning Sample Classification Table
====================================

Actual       Predicted Class                   Actual
Class                     NO          YES       Total
----------------------------------------------------
NO                    313.00        71.00      384.00
```

```
YES                   18.00       235.00       253.00
----------------------------------------------------
Pred. Tot.           331.00       306.00       637.00
Correct             0.81510      0.92885
Success Ind.        0.21228      0.53168
Tot. Correct        0.86028
----------------------------------------------------

================================================
Learning Sample Classification Probability Table
================================================

Actual       Predicted Class                 Actual
Class                    NO          YES      Total
----------------------------------------------------
NO                  0.81510      0.18490    1.00000
YES                 0.07115      0.92885    1.00000
----------------------------------------------------
```

Supporting Information: Appendix 6

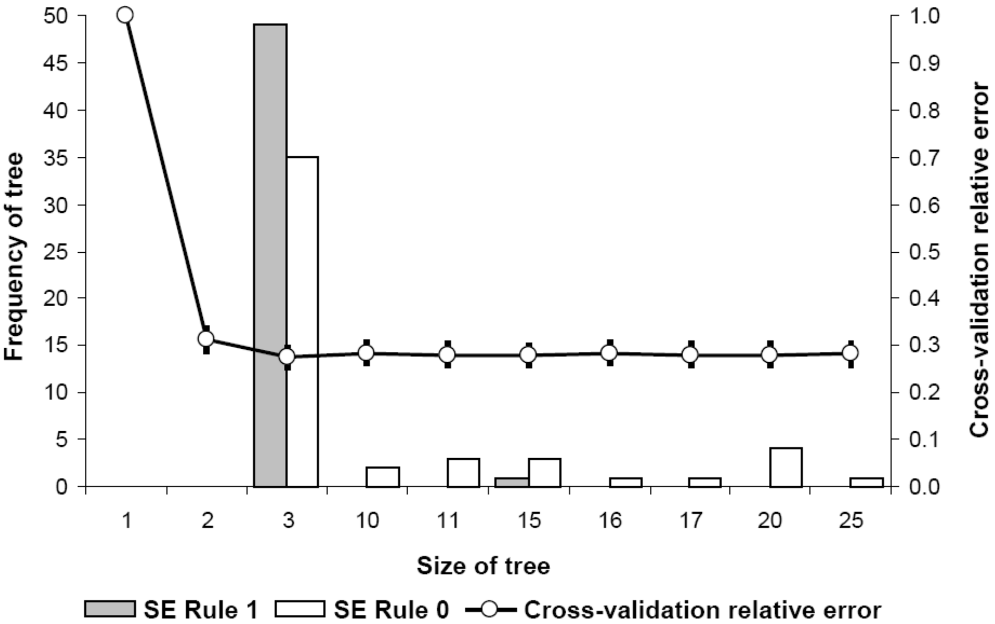


Figure. Cross-validation processes for the selection of the optimal regression tree (Fig. 3) of the probability of non-native species presence in Kruger National Park (line, a single representative 10-fold cross-validation of the most frequent [modal] optimal tree with SE estimate of each tree size; bars, number of optimal trees of each size (frequency of tree) selected from a series of 50 cross-validations based on the minimum cost tree, which minimizes the cross-validated error (SE rule 0), and 50 cross-validations based on the SE rule 1, which minimizes the cross-validated error within 1 SE of the minimum. The most frequent (modal) tree has three terminal nodes.

Supporting Information: Appendix 7

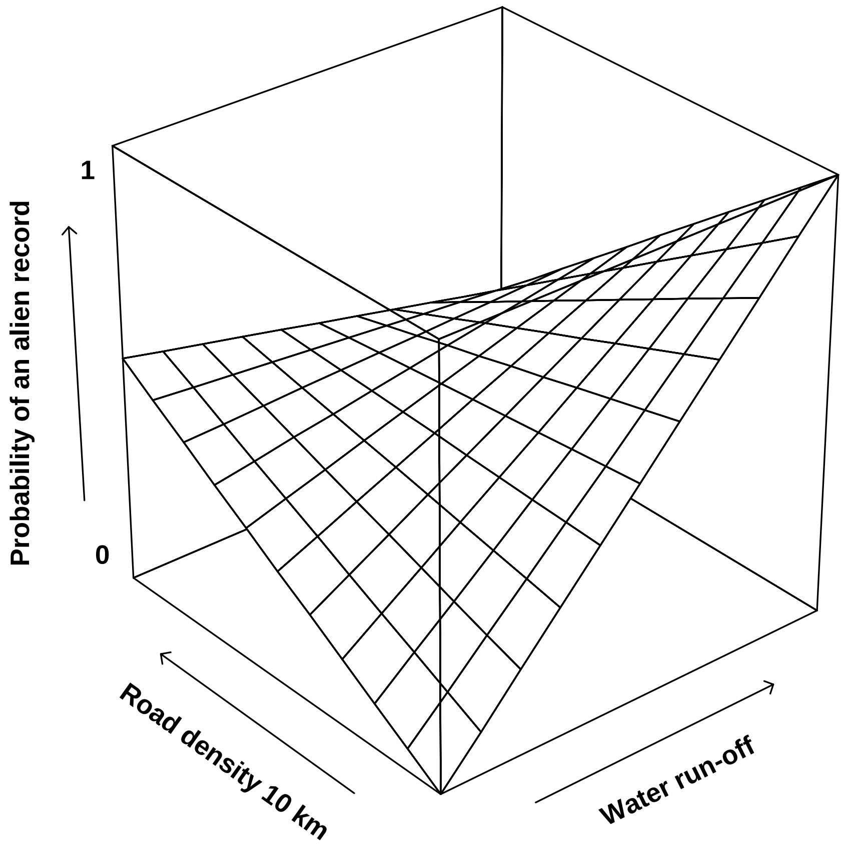


Figure. Logistic model explaining the probability of non-native plant presence in boundary segments. The probability of non-native plant presence was associated with water runoff (0–27 million m^3) and major road density within 10 km radius outside the KNP boundary (0–0.15 km/km^2). Overall significance of the model is $G^2 = 531.54$, df=3, p <0.0001. Parameters of the model are given in Appendix 8.

Supporting information: Appendix 8

Table. Values of the logistic model explaining the presence of non-native plant species in Kruger National Park.[a]

Variable	Estimate	ASE	Standardized estimate	ASE	G^2	df	*p*
Intercept	-6.30	0.66	-0.73	0.18			
Water runoff[b]	0.74	0.08	3.37	0.31	264.82	1	< 0.0001
Road density[c]	69.70	8.59	1.54	0.20	127.96	1	< 0.0001
Runoff × road density	-5.90	0.89	-1.63	0.25	49.13	1	< 0.0001

[a]Standardized estimates comparable at the same scale (ASE, asymptotic standard error of the variable). Results of the G^2 tests describe significance of results with likelihood ratio χ^2 statistics. Overall goodness of fit determined with Hosmer and Lemeshow (1989) test : $\chi^2 = 9.71$, df = 6, NS (spell out); overall explained variance, $r^2_L = 0.66$; correct classification of presence of a non-native species in a segment 90.4%, absence 92.0%, and presence and absence combined 91.3%. Forty segments that caused a lack of fit in the overall goodness of fit of the model [Hosmer and Lemeshow test (1989) for all segments: $\chi^2 = 54.20$, df = 6, $p < 0.0001$]. No autocorrelation coefficients among Pearson's standardized residuals of the model are statistically significant.

[b]Million m^3 / quaternary watershed / annum

[c]Km/km^2

Supporting Information: Appendix 9

Table. Summary and diagnostic information on the classification tree analysis of the probability of presence /absence (yes/no in the terminal node information below) of a non-native plant in Kruger National Park boundary segments, for Fig. 4.

Terminal node information. The first row for each terminal node shows the terminal node number (Node), number of cases at the node (N), fraction of the data that end up in the node (Prob), fraction of records at the node misclassified (Cost), and the dominant splitting category of the response variable for the node (Class). The following rows show distribution of the categories of the response variable at the node (N = numbers, Prob = probabilities of the numbers, Class = category).

```
=========================
TERMINAL NODE INFORMATION
=========================

 Node            N   Prob   Cost  Class
-----------------------------------------
    1          286 0.3785 0.0470  NO

               277 0.9530         NO
                 9 0.0470         YES

    2           28 0.0486 0.2680  YES

                10 0.2680         NO
                18 0.7320         YES

    3          217 0.4059 0.1091  YES

                34 0.1091         NO
               183 0.8909         YES

    4           38 0.0522 0.1515  NO

                34 0.8485         NO
                 4 0.1515         YES

    5           68 0.1148 0.3288  YES

                29 0.3288         NO
                39 0.6712         YES
```

Misclassification by class. Category of the response variable (Class), probabilities of finding a member of each class in the whole sample (Prior Prob.), number of cases in each class (Count), number misclassified (Misclass), and rate of misclassification for each class (Cost). Data set used to grow the tree, i.e. the learning sample, shows the results when these data are dropped down the chosen tree. Data in brackets are based on cross validation, i.e. the best estimates of the misclassification that would occur if the classification tree were to be applied to new data, assuming that the new data were drawn from the same distribution as the learning data.

```
=========================
MISCLASSIFICATION BY CLASS
=========================

(Cross Validation)
              Prior
 Class         Prob        Count     Misclass         Cost
----------------------------------------------------------
NO          0.50000          384         73        0.19010
                            (384         76        0.19792)
YES         0.50000          253         13        0.05138
                            (253         25        0.09881)
----------------------------------------------------------
 Total      1.00000          637        86.00
                            (637       101.00)
```

Prediction success classification and probability tables. Cases of the classification and probability tables appearing on the diagonals of the matrices correspond to correct classification, while off-diagonal entries represent misclassification; the tables make it easy to see where the misclassifications are occurring. The learning sample tables are generated from the application of the input data to the selected optimal tree, and the cross-validation tables constitute the best estimate of the results that would occur if the tree were applied to fresh data. Pred. Tot. = predicted total in each column or class; Correct = percentage classified correctly; Tot. correct = fraction of cases classified correctly overall; Success ind. = fraction by which the percent correct exceeds what would be expected on the basis of chance alone.

```
=====================================
Cross Validation Classification Table
=====================================

Actual       Predicted Class                  Actual
Class                   NO          YES        Total
----------------------------------------------------
NO                  308.00        76.00       384.00
YES                  25.00       228.00       253.00
----------------------------------------------------
Pred. Tot.          333.00       304.00       637.00
Correct            0.80208      0.90119
Success Ind.       0.19926      0.50401
Tot. Correct       0.84144
-----------------------------------------------------------

=================================================
Cross Validation Classification Probability Table
=================================================

Actual       Predicted Class                  Actual
Class                   NO          YES        Total
----------------------------------------------------
NO                 0.80208      0.19792      1.00000
```

```
YES                0.09881      0.90119      1.00000
----------------------------------------------------
====================================
Learning Sample Classification Table
====================================

Actual       Predicted Class               Actual
Class                   NO          YES       Total
----------------------------------------------------
NO                  311.00        73.00      384.00
YES                  13.00       240.00      253.00
----------------------------------------------------
Pred. Tot.          324.00       313.00      637.00
Correct            0.80990      0.94862
Success Ind.       0.20707      0.55144
Tot. Correct       0.86499
-----------------------------------------------------------


================================================
Learning Sample Classification Probability Table
================================================

Actual       Predicted Class               Actual
Class                   NO          YES       Total
----------------------------------------------------
NO                 0.80990      0.19010      1.00000
YES                0.05138      0.94862      1.00000
----------------------------------------------------
```

Supporting Information: Appendix 10

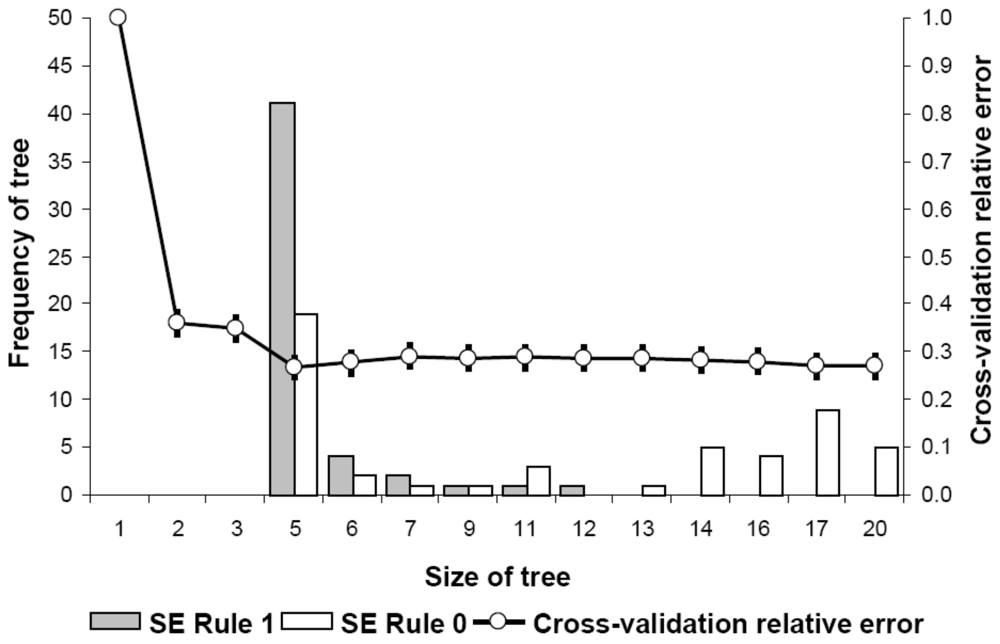


Figure. Graph is cross-validation processes for the selection of the optimal regression tree (Fig. 4). The most frequent (modal) tree has 5 terminal nodes; other elements are as in Fig. 3.